AD-A053 613 OFFICE OF NAVAL RESEARCH ARLINGTON VA F/G 15/5
NAVAL RESEARCH LOGISTICS QUARTERLY. VOLUME 24, NUMBER 4.(U)
DEC 77

UNCLASSIFIED NL

1 OF 2
AD
A053613

NAVAL RESEARCH LOGISTICS QUARTERLY

Volume 24 Number 4.

DEC~~EMBER~~ 1977
VOL. 24, NO. 4

OFFICE OF NAVAL RESEARCH

NAVSO P-1278

NAVAL RESEARCH LOGISTICS QUARTERLY

The Naval Research Logistics Quarterly is devoted to the dissemination of scientific information in logistics and will publish research and expository papers, including those in certain areas of mathematics, statistics, and economics, relevant to the over-all effort to improve the efficiency and effectiveness of logistics operations.

Information for Contributors is indicated on inside back cover.

The Naval Research Logistics Quarterly is published by the Office of Naval Research in the months of March, June, September, and December and can be purchased from the Superintendent of Documents, U.S. Government Printing Office, Washington, D.C. 20402. Subscription Price: $11.15 a year in the U.S. and Canada, $13.95 elsewhere. Cost of individual issues may be obtained from the Superintendent of Documents.

The views and opinions expressed in this Journal are those of the authors and not necessarily those of the Office of Naval Research.

Issuance of this periodical approved in accordance with Department of the Navy Publications and Printing Regulations, P-35 (Revised 1-74).

QUEUEING MODELS FOR SPARES PROVISIONING*

Donald Gross, Henry D. Kahn, Joseph D. Marsh

Department of Operations Research
The George Washington University
Washington, D.C.

ABSTRACT

A population of items which break down at random times and require repair is studied (the classic "machine repair problem with spares"). It is desired to determine the number of repair channels and spares required over a multiyear planning horizon in which population size and component reliability varies, and a service level constraint is imposed. When an item fails, a spare (if available) is immediately dispatched to replace the failed item. The failed item is removed, transported to the repair depot, repaired, and then placed in the spares pool (which is constrained to be empty not more than 10% of the time) unless there is a backlog of requests for spares, in which case it is dispatched immediately. The first model considered treats removal, transportation, and repair as one service operation. The second model is a series queue which allows for the separate treatment of removal, transportation, and repair. Breakdowns are assumed Poisson and repair times exponential.

1. INTRODUCTION

One of the earliest applications of queueing theory to spares provisioning problems was the work of Taylor and Jackson [12], in which a finite source queueing model with spares was used to determine the number of spare engines required to maintain a fleet of aircraft at a certain efficiency level. A bibliography of some recent work on queueing approaches to provisioning type problems is given in Lureau [6]. Of particular interest are Refs. [2], [8], and [9]. The problem treated in this paper is the determination of an adequate number of spares and repair lines (servers) for replacing and repairing components which randomly fail, assuming that the failed components are replaced by spares (if available) and once repaired, in turn become spares. A *multiyear planning horizon* is considered, allowing for growth both in component population size and component reliability.

When a component fails, a request for a spare is immediately placed. If no spare is available, a delay occurs. A service level constraint of 10% is imposed on such delays; that is, we desire a capability such that at least 90% of the requests for spares are immediately filled from on-hand spares inventory (at most, 10% backordering of spares requests). This service criterion is often referred to as availability or fill rate.

***This work was done under the Office of Naval Research Contract N00014-75-C-0729, Project NR 347 020.**

The objective then becomes one of minimizing expenditures for spares and servers subject to a 90% fill-rate constraint. The classical machine repair queueing model with spares is the driving model in that it determines fill rate given a specific number of spares and servers, the component population size, failure rate, and service rate. Times between failures are assumed to be exponentially distributed random variables, as are service times (which include the time required for removal, transportation, and repair). This queueing model is then embedded in a heuristic cost optimization model which determines a "good" mix of spares and servers for each year in the planning horizon while satisfying the fill-rate constraints.

An alternate queueing model is also presented which treats removal, transportation, and repair as separate serving stations, but which must assume an "infinite" calling population. The accuracy of this assumption is also investigated.

The methodology presented here has been used for provisioning servers and spares for a fleet of marine gas turbine engine ships (each engine having two components—a gas generator and a power turbine). It is, of course, applicable to other similar provisioning problems.

2. QUEUEING MODEL

In order to determine the probability of a request for a spare being met without delay, it is necessary to calculate the probability of various numbers of components in the repair system at any particular time. Changes in the population with respect to size and reliability take place at random times throughout each year, so that unless (or until) population size and component reliability stop changing, steady state cannot be approached. It appears analytically intractable to calculate transient state probabilities. As an approximation, we consider the population to be in steady state at its average size and reliability for an entire year, changing instantaneously to a new steady-state position at new average values at the beginning of each new year. In situations where failures are frequent, transient effects should die out quickly and our steady-state approximations should be adequate. However, if failures are infrequent, the transient effects will take longer to disappear and might present a problem. Nevertheless, for most provisioning purposes, the assumption of instantaneous steady-state should be adequate.

At any point in time, the population is composed of units which may have different failure rates, since units added to the population in later years generally are more reliable due to technological learning. It is assumed here that for each year, all components have identical failure rates equal to the population average, which changes on a yearly basis as described above. This assumption that all components operate at the population average failure rate was investigated in Ref. [5], and the results will be briefly described later in this paper.

We introduce the following notation:

$p_{n,i}$=steady-state Pr {n components in repair, year i}
c_i=number of repair facilities, year i
y_i=number of spares, year i
λ_i=component failure rate (Poisson mean), year i
$\bar{\lambda}_i$=population average failure rate, year i
$\bar{R}_i$=expected number of components repaired, year i
N_i=component population size, year i
$1/\mu_i$=average service (removal, transportation, and repair) time, (exponential mean), year i.

The equations for the steady-state probability of n items "down" (i.e., in repair) are given as

$$(1)\qquad p_{n,i}=\begin{cases}\begin{cases}\dfrac{N_i^n}{n!}\left(\dfrac{\bar{\lambda}_i}{\mu_i}\right)^n p_{0,i} & (0\leq n\leq c_i)\\[2ex] \dfrac{N_i^n}{c_i^{n-c_i}c_i!}\left(\dfrac{\bar{\lambda}_i}{\mu_i}\right)^n p_{0,i} & (c_i\leq n\leq y_i)\\[2ex] \dfrac{N_i^{y_i}N_i!}{(N_i-n+y_i)!\,c_i^{n-c_i}c_i!}\left(\dfrac{\bar{\lambda}_i}{\mu_i}\right)^n p_{0,i} & (y_i<n\leq y_i+N_i)\end{cases} & \langle c_i\leq y_i\rangle\\[8ex] \begin{cases}\dfrac{N_i^n}{n!}\left(\dfrac{\bar{\lambda}_i}{\mu_i}\right)^n p_{0,i} & (0\leq n\leq y_i)\\[2ex] \dfrac{N_i^{y_i}N_i!}{(N_i-n+y_i)!\,n!}\left(\dfrac{\bar{\lambda}_i}{\mu_i}\right)^n p_{0,i} & (y_i<n\leq c_i)\\[2ex] \dfrac{N_i^{y_i}N_i!}{(N_i-n+y_i)!\,c_i^{n-c_i}c_i!}\left(\dfrac{\bar{\lambda}_i}{\mu_i}\right)^n p_{0,i} & (c_i<n\leq y_i+N_i)\end{cases} & \langle c_i>y_i\rangle\end{cases}$$

$$p_{0,i}:\sum_{n=0}^{N_i+y_i} p_{n,i}=1$$

$$(2)\qquad \bar{\lambda}_i=\begin{cases}\{(N_i-N_{i-1})\lambda_i+\bar{R}_{i-1}\lambda_{i-1}+(N_{i-1}-\bar{R}_{i-1})\bar{\lambda}_{i-1}\}/N_i, & N_i\geq N_{i-1}\\ \{\bar{R}_{i-1}\lambda_{i-1}+(N_{i-1}-\bar{R}_{i-1})\bar{\lambda}_{i-1}\}/N_{i-1}, & N_i\leq N_{i-1}\end{cases}$$

$$(3)\qquad \bar{R}_i=\bar{\lambda}_i\left[N_i-\sum_{n=y_i+1}^{N_i+y_i}(n-y_i)p_{n,i}\right].$$

Equation Set (1) gives the standard fomulas for a machine repair model with spares (see Gross and Harris [4], p. 123, for example).

Using the average failure rate $\bar{\lambda}$ for each year allows for the incorporation of changing component reliability as the years progress. Equation Set (2) shows how $\bar{\lambda}$ is computed. It is, perhaps, easiest to explain in the context of the marine gas turbine engine example mentioned previously. Year one begins with the first introduction of a few gas turbine ships, and each succeeding year brings into the fleet additional ships, until the fleet reaches full strength. The components introduced in the later years generally have improved reliability, either through learning or a conscious component improvement program (CIP). Thus, λ_i tends to be smaller than λ_{i-1}. When the population size is increasing ($N_i>N_{i-1}$), the average failure rate over all components in the population for year i, $\bar{\lambda}_i$, is given as the weighted average of the new components introduced into the population at the best current achievable failure rate λ_i, those repaired during the past year at that year's best achievable failure rate λ_{i-1}, and those old components in the population which were not repaired and are thus operating at the old average failure rate $\bar{\lambda}_{i-1}$. If the fleet size is decreasing ($N_i<N_{i-1}$), (i.e., some ships may be retired from service), then the computation is given by the second equation of (2), namely, the weighted average of those repaired last year coming in with a failure rate of λ_{i-1} and those not repaired but operating at the old average $\bar{\lambda}_{i-1}$. The average number repaired in a year is given by (3).

The assumption that all components operate at an average failure rate was investigated in Ref. [5] by developing the exact model for unequal failure rates with $N=1$, $Y=1$, and $c=1$, so that one component has a failure rate of λ_1 while the other has a failure rate of λ_2. It turns out that (1), assuming both units operate at $\bar{\lambda}=(\lambda_1+\lambda_2)/2$, gives conservative results with respect to availability; i.e., the actual availability is greater than that computed using $\bar{\lambda}$, and (2) even for sizable differences in λ_1 and λ_2, the approximate availability using $\bar{\lambda}$ is close to the actual. The reader is referred to Ref. [5] for more detail. Result (1) has some intuitive appeal, since components having higher failure rates would be expected to fail (and hence be in repair) more often. Thus, the more reliable components operate a larger proportion of the time, so that the actual failure rate for the system would tend to be lower than the arithmetic average of the failure rates over all components. Nevertheless, it appears from Result (2) that this effect is not highly significant and that using the population average failure rate for each component is an adequate approximation in most cases. For example, using the exact model with $N=Y=c=1$, it was found that, even for λ_2 twice that of λ_1, the percentage error in availability was no more than 5% when using $\bar{\lambda}$ as the failure rate for each component. In actual applications, the individual component failure rates tend to be much closer together than a factor of two, since reliability growth is gradual, and hence we feel confident that the average failure rate assumption is justifiable.

Once the $p_{n,i}$ are obtained, the fill-rate constraint is given (temporarily dropping the subscript i) by

$$\sum_{n=0}^{y-1} p_n \geq 0.90,$$

since when n components are down, there are $y-n$ spares available ($n<y$). Thus, the probability of no spares on hand, P_{out}, is

$$1-\sum_{n=0}^{y-1} p_n$$

and the percentage of requests filled immediately from on-hand spares is

$$\lambda(1-P_{\text{out}})/\lambda=1-P_{\text{out}}=\sum_{n=0}^{y-1} p_n.$$

However, the probability that a *failed component finds n in the system* should be used in the fill-rate computation. In common queueing terminology, these are the "arriving customer" probabilities and, in the context of this paper, correspond to the occurrence of component failures which generate requests for spares. Thus, they shall be referred to as failure point probabilities. For the finite source with spares queue considered in this paper, the failure point probabilities are not equal to the general time probabilities given by (1). For the finite source-no spares queue (see, e.g. Cooper [1], pp. 82 ff.), the failure point probabilities are equivalent to the general time probabilities for a finite calling population of one less, but this relationship does not hold for the spares case.

The failure point probabilities for the finite source with spares queue (denoted by q_n, as contrasted to the general time probabilities, denoted by p_n) may be derived as follows. Using Bayes' theorem,

$$q_n \equiv \Pr\{n \text{ in system} \mid \text{failure about to occur}\}$$

$$=\frac{\Pr\{n \text{ in system}\}\ \Pr\{\text{failure about to occur} \mid n \text{ in system}\}}{\sum_n [\Pr\{n \text{ in system}\}\ \Pr\{\text{failure about to occur} \mid n \text{ in system}\}]}$$

Now, since this is a birth-death process with

$$\Pr\{\text{failure in } t,\, t+\Delta t\}=\lambda_n\Delta t+o(\Delta t),$$

where

$$\lambda_n=\begin{cases} N\lambda, & (0\leq n<y)\\ (N-n+y)\lambda, & (y\leq n\leq y+N),\\ 0, & (n>y+N)\end{cases}$$

we obtain

$$q_n=\begin{cases} \lim\limits_{\Delta t\to 0}\dfrac{p_n[N\lambda\Delta t+o(\Delta t)]}{\sum\limits_{n=0}^{y-1}p_n[N\lambda\Delta t+o(\Delta t)]+\sum\limits_{n=y}^{y+N}p_n[(N-n+y)\lambda\Delta t+o(\Delta t)]}, & (0\leq n<y),\\[2ex] \lim\limits_{\Delta t\to 0}\dfrac{p_n[(N-n+y)\lambda\Delta t+o(\Delta t)]}{\sum\limits_{n=0}^{y-1}p_n[N\lambda\Delta t+o(\Delta t)]+\sum\limits_{n=y}^{y+N}p_n[(N-n+y)\lambda\Delta t+o(\Delta t)]}, & (y\leq n\leq y+N).\end{cases}$$

Dividing numerator and denominator by Δt and taking the limit yields

$$q_n=\begin{cases} \dfrac{Np_n}{N-\sum\limits_{n=y}^{y+N}(n-y)p_n}, & (0\leq n<y),\\[2ex] \dfrac{(N-n+y)p_n}{N-\sum\limits_{n=y}^{y+N}(n-y)p_n}, & (y\leq n\leq y+N),\end{cases}$$

where the p_n are given in Equation (1). Thus, the fill-rate constraints must be based on the q_n, that is

$$\sum_{n=0}^{y-1}q_n\geq 0.90.$$

For the $\bar{R}$ calculation, the general time probability is required so that (1) is used as given, and, in fact, $\bar{R}$ equals the denominator term of q_n multiplied by λ.

3. COST MODEL

The cost model considers four types of costs: (a) purchase cost of spares, (b) purchase cost of service channels, (c) repair costs, and (d) investment costs in component improvement (reliability improvement) programs. Annual operating costs associated with running the spares inventory system or operating the service channels are not included. While the variable portion of these costs can be important, they are not considered in this model, since the major purpose of such a strategic planning model as this is for capital budgeting and, hence, it is the purchase expenditures which are of prime concern at this stage of the decision-making process.

The optimization problem is an integer-nonlinear-programming problem with a hard-to-manage objective function and highly complex constraints. An integer programming algorithm approach has not as yet been successful in solving the problem, hence, a heuristic method is utilized.

The heuristic cost "optimization" algorithm considers explicitly only the first two categories of costs, although the repair and component improvement program (CIP) costs are always calculated for each year so that sensitivity analyses with respect to various CIP programs, service times, etc., can be performed. It should be noted that even with no CIP program, $\bar{\lambda}$ and N would still change, but the CIP program influences the magnitudes of these changes. The present worth of the cost stream over the planning horizon is always calculated for just such a use. This will be illustrated in the next section.

The cost minimization problem can be stated as follows. Considering the additional notation

$C_{1,i}$=purchase cost of a repair channel, year i
$C_{2,i}$=purchase cost of a spare component, year i
$C_{3,i}$=cost of repairing a component, year i
$C_{4,i}$=investment in reliability growth (CIP) program, year i
d=discount factor
k=number of years in planning horizon,

we desire to

$$\min_{c_1, c_2, \ldots, c_k;\, y_1, y_2, \ldots, y_k} Z=\sum_{i=1}^{k} d^i[C_{1,i}(c_i-c_{i-1})^+ + C_{2,i}(y_i-y_{i-1})^+ + C_{3,i}\bar{R}_i + C_{4,i}] \tag{4}$$

$$\text{Subject to: } \sum_{n=0}^{y-1} q_{n,i} \geq 0.90, \qquad (i=1, 2 \ldots, k) \tag{5}$$

$$c_i \geq 0;\ \text{integer.}$$

$$y_i \geq 0;\ \text{integer.}$$

The plus superscript on the first two cost terms indicates that the term is zero if the factor in parentheses is negative; that is,

$$(a-b)^+ = \max\{(a-b), 0\}.$$

The last term in Equation (4), $C_{4,i}$, is the cost of the CIP and does not explicitly depend on c_i or y_i, although indirectly one can argue that if $C_{4,i}$ is increased, λ_i will be smaller, thus affecting the $p_{n,i}$ calculation and hence the constraint (5). The heuristic algorithm ignores this. This CIP effect is observed via a sensitivity type of analysis referred to above.

The third term, which involves $\bar{R}_i$, is a function of y_i (explicitly, see Equation (3)), and y_i and c_i implicitly through $p_{n,i}$. However, when compared to purchase costs of spares and servers, the annual repair costs should be small and are ignored by the heuristic algorithm. It is included in the cost calculation because it is directly influenced by reliability growth and is needed for any sensitivity study of CIP. Thus, the heuristic algorithm looks only at the costs of purchasing spares and servers and, furthermore, considers only one year at a time; that is, it attempts to

$$\min_{c_i, y_i} Z_i = C_{1,i}(c_i-c_{i-1})^+ + C_{2,i}(y_i-y_{i-1})^+$$

for each $i=1, 2, \ldots, k$ in a sequential manner, using the "best" combination it finds for year i-1 (c_{i-1}, y_{i-1}) as the starting point for its search for (c_i, y_i).

For the year i computations, the algorithm first checks to see if c_{i-1}, y_{i-1} satisfy the constraint. If the constraint is not met, the algorithm computes the ratio $\Delta=C_{2,i}/C_{1,i}$ and takes the largest integer value contained therein. (In the applications we have studied, $C_{2,i}$ was always larger than $C_{1,i}$. If this is not the case, the algorithm should be modified to compute $C_{1,i}/C_{2,i}$, and the words "server" and "spare" interchanged in the description that follows.) Thus, for an (approximately) equal dollar expenditure, we can purchase one spare or Δ servers. Both points $(c_{i-1}+\Delta, y_{i-1})$ and $(c_{i-1}, y_{i-1}+1)$ are checked to determine the fill-rate percentage via use of the queueing model, and the point with the largest fill-rate percentage is chosen. This procedure is continued from the "new points" chosen until the fill rate reaches (or exceeds) 90%. If 90% is exceeded (and it generally will be because of the integer values required for c and y), a reduction procedure is employed to see if any servers can be decreased (since Δ servers at a time have been added) and still maintain a 90% fill rate. Also, if the feasible region has been entered by adding Δ servers (as opposed to adding a spare) in addition to decreasing servers, an attempt is made to reduce the number of spares. This reduction portion of the algorithm is intended to find a solution as near to the constraint boundary as possible. The procedure is schematically shown in Figure 1, which displays all the points (c, y) evaluated by the algorithm along with the path of greatest improvement in fill rate. Note that for each point on the path, fill rate is calculated for the points immediately above and to the right, and then a move is made to the one with the largest fill rate.

Should the starting value for year $i\,(c_{i-1}, y_{i-1})$ be such that the 90% fill rate criterion is exceeded (this may happen if population size is decreasing, i.e., units are being retired from service), then the algorithm immediately goes into the reduction mode, first trying to reduce, to the greatest extent possible, spares (assuming that they are more expensive than servers and hence should have a larger salvage value) and then attempting to reduce servers to get as close to the 90% boundary line as possible. Again, if servers are more expensive than spares, the words "server" and "spare" should be interchanged.

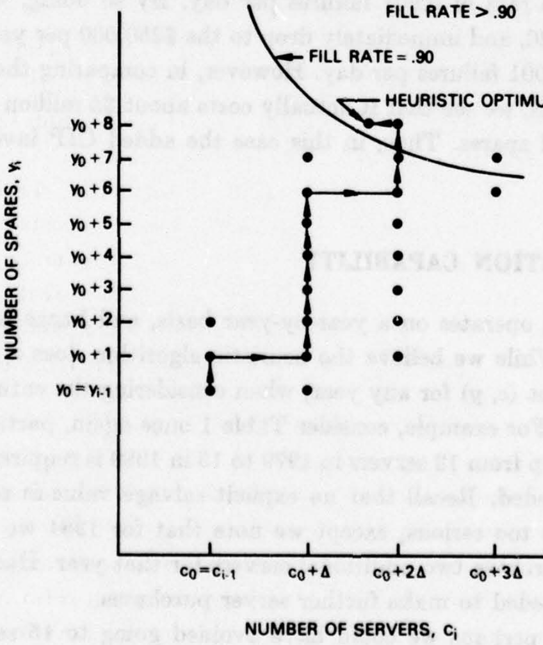


FIGURE 1. Heuristic algorithm.

4. SAMPLE RESULTS

Using the marine gas turbine fleet as an example, Table 1 shows the results for an 11-year planning horizon, from 1975 to 1985, using an interest rate of 10%. The fleet size starts out at a relatively small number, building up to full strength by year 1985. The component exhibited is the gas generator component. Similar calculations were performed for the power turbine component. Generators and turbines are completely independent as they require different repair facilities, and, of course, spares are not interchangeable.

The first nine columns of Table 1 (except the $\bar{\lambda}$ column, column 4) are input. Column 1 gives the year; Column 2, the anticipated population size; Column 3, the mean failure rate schedule for the particular CIP chosen (in failures per day); Column 5, the average service (including removal, transportation, and repair) time (in days); and Columns 6–9, the purchase cost of servers, purchase cost of spares, unit repair cost, and investment in CIP cost, respectively, with C_1 through C_3 being in thousands of dollars per unit, while C_4 is in thousands of dollars per year.

Column 4 and the last five columns are output. It was convenient to show $\bar{\lambda}_t$ (Column 4) next to λ_t so that one can readily see the reliability growth. Columns 10 and 11 give the algorithm's "best" c_t, y_t. Column 12 shows average number of units repaired for each year, while Column 13 provides the expenditures (in thousands of dollars) for year i. Column 14 gives the present worth of the discounted cost stream up to and including year i. The final value in Column 14 is the present worth of all expenditures over the complete planning horizon, again in thousands of dollars.

Another run is presented in Table 2 which differs only in the CIP chosen. This illustrates how one can study the consequence of various alternative CIP's. There we assume that in order to save CIP investment cost (C_4), we will stop the program at year 1978, thus achieving only a minimum component failure rate of 0.001 failures per day. By so doing, we save the two large expenditures in 1979 and 1980, and immediately drop to the $350,000 per year needed to maintain the achieved reliability of 0.001 failures per day. However, in comparing the present worth of the cost streams for the two cases, we see that it actually costs about $5 million more to do this, since we require more servers and spares. Thus, in this case the added CIP investment is well worth the expenditure.

5. PROGRAM PERTURBATION CAPABILITY

The heuristic algorithm operates on a year-by-year basis, and hence has the limitation that it does not "look ahead." While we believe the heuristic algorithm does a good job in giving an optimal or near optimal point (c, y) for any year, when considering the entire planning horizon, it may not prove as effective. For example, consider Table 1 once again, particularly year 1980. We see that, for that year, a jump from 12 servers in 1979 to 15 in 1980 is required, but in the following year only 13 servers are needed. Recall that no explicit salvage value is received for decreasing servers. This in itself is not too serious, except we note that for 1984 we must increase servers back up to 15, having to purchase two additional servers for that year. Had we kept the 15 from 1981, we would not have needed to make further server purchases.

But more importantly, perhaps we could have avoided going to 15 servers in 1980 had we purchased an additional spare, especially since it was needed for year 1981 anyway. The algorithm allows us to go back to 1980 and use as a starting point $c=12$, $y=13$ (instead of the 1979 solution

TABLE 1. *Gas Generator—Full* CIP

Year	N	λ	$\overline{\lambda}$	$1/\mu$	C_1	C_2	C_3	C_4	c	y	$\overline{R}$	Cost	Present Worth
75	10	0. 0015	0. 0015	65. 0	132. 0	822. 0	49. 0	1975. 0	3	3	5. 3	5097. 61	5097. 61
76	28	0. 0015	0. 0015	62. 5	132. 0	945. 0	49. 0	2760. 0	5	6	15. 3	6607. 21	11104. 17
77	50	0. 0014	0. 0015	60. 0	132. 0	1087. 0	37. 8	3840. 0	8	8	26. 4	7407. 54	17226. 10
78	82	0. 0010	0. 0013	57. 5	132. 0	1174. 0	40. 0	3950. 0	10	10	37. 2	8048. 63	23273. 15
79	121	0. 0008	0. 0010	55. 0	132. 0	1268. 0	42. 0	2800. 0	12	11	45. 5	6243. 08	27537. 26
80	158	0. 0007	0. 0009	55. 0	132. 0	1369. 0	44. 0	1100. 0	15	12	51. 8	5144. 28	30731. 45
81	182	0. 0006	0. 0008	55. 0	132. 0	1369. 0	44. 0	350. 0	13	13	53. 9	4092. 71	33041. 68
82	208	0. 0006	0. 0007	55. 0	132. 0	1369. 0	44. 0	350. 0	14	13	56. 3	2957. 99	34559. 60
83	229	0. 0006	0. 0007	55. 0	132. 0	1369. 0	44. 0	350. 0	13	14	58. 3	4282. 26	36557. 30
84	251	0. 0006	0. 0006	55. 0	132. 0	1369. 0	44. 0	350. 0	15	14	61. 6	3323. 18	37966. 66
85	256	0. 0006	0. 0006	55. 0	132. 0	1369. 0	44. 0	350. 0	15	14	61. 6	3059. 91	39146. 38

TABLE 2. *Gas Generator—Partial* CIP

Year	N	λ	$\bar{\lambda}$	$1/\mu$	C_1	C_2	C_3	C_4'	c	y	$\bar{R}$	Cost	Present Worth
75	10	0. 0015	0. 0015	65. 0	132. 0	822. 0	49. 0	1975. 0	3	3	5. 3	5097. 61	5097. 61
76	28	0. 0015	0. 0015	62. 5	132. 0	945. 0	49. 0	2760. 0	5	6	15. 3	6607. 21	11104. 17
77	50	0. 0014	0. 0015	60. 0	132. 0	1087. 0	37. 8	3840. 0	8	8	26. 4	7407. 54	17226. 10
78	82	0. 0010	0. 0013	57. 5	132. 0	1174. 0	40. 0	3950. 0	10	10	37. 2	8048. 63	23272. 15
79	121	0. 0010	0. 0010	55. 0	132. 0	1268. 0	42. 0	350. 0	11	12	47. 6	5016. 79	2669. 69
80	158	0. 0010	0. 0010	55. 0	132. 0	1369. 0	44. 0	350. 0	13	14	59. 0	5947. 59	30392. 67
81	182	0. 0010	0. 0010	55. 0	132. 0	1369. 0	44. 0	350. 0	16	15	66. 7	5049. 24	33242. 83
82	208	0. 0010	0. 0010	55. 0	132. 0	1369. 0	44. 0	350. 0	17	17	75. 0	6548. 08	36603. 03
83	229	0. 0010	0. 0010	55. 0	132. 0	1369. 0	44. 0	350. 0	19	18	83. 4	5651. 54	39239. 51
84	251	0. 0010	0. 0010	55. 0	132. 0	1369. 0	44. 0	350. 0	19	20	91. 9	7129. 79	42263. 24
85	256	0. 0010	0. 0010	55. 0	132. 0	1369. 0	44. 0	350. 0	21	20	94. 3	4761. 92	44099. 16

$c=12$, $y=11$). This may give a different c, y schedule from that year on, which might have a cheaper fixed present worth. This capability to go back to any year in the schedule and use a starting point other than the previous year's c, y we refer to as the perturbation capability. Using a different starting point (if chosen "intelligently") might get us to a different feasible solution point which avoids temporary increases in either c or y. The program then continues its calculation from the year for which the perturbation is desired, adding the new cost stream with the correct discount factor from that year forward, giving us the present worth for the new schedule which may result. Table 3 shows this effect for perturbing the Table 1 solution at year 1980, using the starting point $c=12$, $y=13$ rather than the 1979 solution of $c=12$, $y=11$. We see for the new resulting schedule (which differs only for year 1980) that the total present worth is now about $39.05 million, a savings of approximately $0.1 million. Further possible perturbations are also shown with the resulting costs. No other obvious perturbations are indicated, although earlier purchasing to avoid inflation might be evaluated.

At first glance, it appears this multiyear programming problem might be amenable to a dynamic programming solution where the years are stages and c and y are the state variables. Aside from the fact that there are two state variables (which makes the computations formidable), the problem is not decomposable. The fill-rate constraint must hold for each year. This constraint (e.g., for year i) involves the $q_{n,i}$ which are functions of the $p_{n,i}$ which in turn are functions not only of the decision variables for year i (c_i, y_i), but also of all previous c's and y's, since $\bar{\lambda}_i$ is a function of $\bar{\lambda}_{i-1}$ which is a

TABLE 3. *Perturbed Solutions, Gas Generator—Full* CIP

Year	Original Solution		First Perturbation (1980)		Second Perturbations (1982)		(1983)	
	c	y	c	y	c	y	c	y
1975	3	3	3	3	3	3	3	3
1976	5	6	5	6	5	6	5	6
1977	8	8	8	8	8	8	8	8
1978	10	10	10	10	10	10	10	10
1979	12	11	12	11	12	11	12	11
1980	15	12	12	13	12	13	12	13
1981	13	13	13	13	13	13	13	13
1982	14	13	14	13	13	14	14	13
1983	13	14	13	14	13	14	14	14
1984	15	14	15	14	15	14	15	14
1985	15	14	15	14	15	14	15	14
Present Worth	$39, 146, 380		$39, 051, 460		$39, 046, 080		$38, 995, 440	

function of $\bar{\lambda}_{i-2}$, etc. (see (2)). This prevents starting at the last stage and proceeding backward. Only if the reliability were constant throughout the multiyear planning horizon could dynamic programming be attempted. However, we believe that the heuristic algorithm, coupled with the perturbation capability, does provide good, although not necessarily optimal, schedules.

6. SERIES QUEUEING MODEL

The queueing model portion of the program given by Equation (1) assumes that removal, transportation, and repair are a single service action, so that if all servers are busy and a component fails, it must wait in a queue prior to removal for an available server. Once removal is initiated, no further queues for transportation or repair are encountered.

In many situations, the removal, transportation, and repair phases may be separate, each having their own "servers" with their own queues. To model this situation would require the $p_{n,i}$ and $q_{n,i}$ to be determined by a closed network or cyclic queueing model which can account for the finiteness of the population, but this requires quite involved calculations (see, for example, refs. [3], [10], and [11]). The question then naturally arises as to how crucial the finite source assumption might be, since rather simple queueing formulas exist for infinite source series queues with exponential arrival and service patterns. For example, suppose for component removal, c_1 removal teams are available for the system (c_1 could equal the population size N, in which case we have an ample server model). Assuming that removal times are exponential and failures are Poisson, we can model the removal portion by an $M/M/c_1$ queue. The output of such a queue (assuming an infinite population) is Poisson. Further, suppose there were c_2 transport vehicles available to the system for shipping components to the repair depot. The transportation phase would then be an $M/M/c_2$ queue if we assume that transport times are exponential. Finally, if we assume that there are c_3 repair channels (again with an infinite population), the repair portion becomes an $M/M/c_3$ queue.

If we let $p_{n,i}^{(j)}$ be the probability of n components at the jth phase ($j=1, 2, 3$) for year i, it can be shown (see Gross and Harris [4], p. 203) that the joint probability of l in the removal portion, m in the transportation portion, and n in repair for year i (call $p_{l,m,n;i}$) is given by

$$p_{l,m,n;i}=p_{l,i}^{(1)}p_{m,i}^{(2)}p_{n,i}^{(3)}.$$

If the year subscript i is dropped for the time being, the fill-rate constraint for each year corresponding to Equation (5) becomes

$$\sum_{l}\sum_{m}\sum_{n}p_{l}^{(1)}p_{m}^{(2)}p_{n}^{(3)}\geq 0.90,$$

where the summation is taken over all combinations of l, m, n such that $(l+m+n\leq y-1)$. Here the $q_k^{(j)}$ equal the $p_k^{(j)}$, as we are assuming an infinite source model. Since $p_k^{(j)}$ is readily computable for $M/M/c$ models, a series queue representation is possible as long as we can justify approximating the finite source situation by an infinite source model. If we assume for the moment that we can (this will be discussed more fully below), then the previous methodology can be used if we substitute for the previously used $q_{n,i}$ and $p_{n,i}$ the sum of probabilities

(6) $$\sum_{l+m+k=n}p_{l,i}^{(1)}p_{m,i}^{(2)}p_{k,i}^{(3)},$$

with the arrival rate for this infinite source series queueing model adjusted to be

$$\lambda_i = N_i \bar{\lambda}_i. \tag{7}$$

Once again, if we drop the year subscript i for convenience, the standard $M/M/c$ formulas (see Gross and Harris [4], p. 96, for example) give

$$p_k^{(j)} = \begin{cases} \dfrac{\lambda^k}{k!\,\mu_j{}^k}\,p_0, & (1 \le k \le c_j) \\[2ex] \dfrac{\lambda^k}{c_j^{k-c_j}\,c_j!\,\mu_j{}^k}\,p_0, & (k \ge c_j) \end{cases} \tag{8}$$

$$p_0^{(j)} = \left[\sum_{k=0}^{c_j-1} \frac{1}{k!}\left(\frac{\lambda}{\mu_j}\right)^k + \frac{1}{c_1!}\left(\frac{\lambda}{\mu_j}\right)^{c_j}\left(\frac{c_j\mu_j}{c_j\mu_j-\lambda}\right)\right]^{-1}, \tag{9}$$

with λ being given by (7), c_j being the number of servers for phase j, and μ_j being the mean service rate for phase j ($j=1=$removal, $j=2=$transportation, $j=3=$repair). In some situations, there may even be a fourth phase consisting of transportation from repair depot to spares pool.

The mathematical programming problem which we attacked via a heuristic algorithm is now considerably more complex if c_1 and c_2 are also decision variables, since we now seek, for each year, the best combination (c_1, c_2, c_3, y) and not merely (c, y). We no longer have a single Δ, but three Δ's, say Δ_1, Δ_2, and Δ_3. For equal dollar expenditures then, we can buy one spare or Δ_1 removal teams, or Δ_2 transport vehicles, or Δ_3 repairmen. Further, there are combinations such as $\alpha\Delta_1$ removal teams$+\beta\Delta_2$ transport vehicles$+\gamma\Delta_3$ repairmen $(0 \le \alpha, \beta, \gamma \le 1)$ which might equal the purchase cost of a spare. Conceptually, the extension to our heuristic algorithm would be, given a particular c_1, c_2, c_3, and y, to calculate the increase in fill rate when moving to all points of equal expenditure; that is, to calculate the fill rate for $(c_1, c_2, c_3, y+1)$, $(c_1+\Delta_1, c_2, c_3, y)$, $(c_1, c_2+\Delta_2, c_3, y)$, $(c_1, c_2, c_3+\Delta_3, y)$, and $(c_1+\alpha\Delta_1, c_2+\beta\Delta_2, c_3+\gamma\Delta_3, y)$ for all appropriate α, β, γ. For simplification, the equal expenditure points involving combinations of c_1, c_2, c_3 (those points with the α, β, and γ terms) might be ignored so that at each step we add either one spare, or Δ_1 removal teams, or Δ_2 transports, or Δ_3 repairmen.

If, on the other hand, c_1 and c_2 are not decision variables (and in our applications they were not), then the algorithm need not be modified since we seek only "optimal points" (c_3, y). Thus, once the p_n's and q_n's are replaced by their series model counterparts as given in Equations (6), (8), and (9), the algorithm is exercised as before, with c_1 and c_2 being fixed values. In the next section, we present some results for such a case after first considering the question of accuracy when using an infinite source approximation to a finite population.

7. ACCURACY OF INFINITE SOURCE APPROXIMATIONS

In order to use a series queueing model, the effect of assuming an infinite population for calculating state probabilities when in actuality the population is finite must be investigated. Let us consider now the nonseries finite source repairman with spares model given by Equation (1) and its infinite population counterpart, an $M/M/c$ model with an arrival rate adjusted as in Equation (7), that is, $\lambda = N\bar{\lambda}$. In the $M/M/c$ model, λ is always $N\bar{\lambda}$, regardless of the number of units down.

In the finite source model, λ is state dependent and, in fact, is given in Section 2 for the q_n development as

$$\lambda_n=\begin{cases} N\bar{\lambda}, & (0\leq n\leq y) \\ (N-n+y)\bar{\lambda}, & (y<n\leq y+N) \\ 0, & (n>y+N), \end{cases}$$

where n is the number of units "down," that is, in "repair." Now the overall average arrival rate of the finite source model would be

$$\begin{aligned}\bar{\bar{\lambda}}&=\sum_{n=0}^{\infty}\lambda_n p_n=\sum_{n=0}^{y}N\bar{\lambda}p_n+\sum_{n=y+1}^{y+N}(N-n+y)\bar{\lambda}p_n\\ &=\sum_{n=0}^{y+N}N\bar{\lambda}p_n-\sum_{n=y+1}^{y+N}(n-y)\bar{\lambda}p_n\\ &=N\bar{\lambda}-\bar{\lambda}\sum_{n=y+1}^{y+N}(n-y)p_n=\bar{\lambda}\left[N-\sum_{n=y+1}^{y+N}(n-y)p_n\right],\end{aligned}$$

which incidentally is how $\bar{R}$ of (3) is derived. Thus, if

$$\sum_{n=y+1}^{y+N}(n-y)p_n$$

is small in relation to N, the infinite source approximation should be adequate, since $\bar{\bar{\lambda}}\doteq N\bar{\lambda}$, the quantity assumed by the infinite source model. This should be the case for systems with not too much congestion (p_n small for $n>y$), y and N large. The fact that a fill rate of 90% must be guaranteed should require that c and y be large enough such that there is a relatively small amount of congestion in the system, thus making the approximation good even for small N.

Table 4, first and third blocks, show comparisons for the two cases run in Tables 1 and 2, respectively, with an infinite source $M/M/c$ used in calculating $p_{n,t}$. The results differ only slightly, with the infinite source model requiring an additional server or two in a few of the years. The infinite source model will occasionally require slightly higher c (or y) than the finite source model in order to meet the availability constraint. This is caused by the lack of state dependence in the arrival rate for the infinite source model, which has the effect of making the infinite source arrival rate slightly higher than the arrival rate for the corresponding finite source model. Hence, the infinite source model is a conservative approximation with respect to fill rate.

Since the infinite source approximation appears to give good results, we have run an infinite source three-stage (removal, transport, and repair) series model, using the same data as in Tables 1 and 2, but assuming that removal is $M/M/\infty$, transport is $M/M/\infty$, and repair is $M/M/c$ (c to be determined as before).

Actually, for ample server queues, it is not necessary to assume exponential service since the output of $M/G/\infty$ is also Poisson (see, for example, Mirasol [7]). Thus, these results also hold for $M/G/\infty$ removal and transport phases. The mean single-stage repair time is allocated to the multistage model so that, on the average, 5% is consumed in removal, 20% in transportation, and 75% in repair. These results comparing the two models are also shown in Table 4. Here, in many of the

TABLE 4. *Comparison of Finite Source, Infinite Source, and Series Models*

	Case 1—Full Component Improvement Program						Case 2—Partial Component Improvement Program					
Year	Single-Stage Models				Multistage Model		Single-Stage Models				Multistage Model	
	Finite Source		$M/M/c$		$M/G/\infty \rightarrow M/G/\infty \rightarrow M/M/c$		Finite Source		$M/M/c$		$M/G/\infty \rightarrow M/G/\infty \rightarrow M/M/c$	
	c	y	c	y	c	y	c	y	c	y	c	y
1975	3	3	3	3	3	3	3	3	3	3	3	3
1976	5	6	5	6	4	6	5	6	5	6	4	6
1977	8	8	8	8	7	8	8	8	8	8	7	8
1978	10	10	10	10	8	10	10	10	10	10	8	10
1979	12	11	12	11	10	11	11	12	11	12	9	12
1980	15	12	16	12	13	12	13	14	14	14	11	14
1981	13	13	13	13	10	13	16	15	16	15	13	15
1982	14	13	14	13	12	13	17	17	17	17	14	17
1983	13	14	13	14	11	14	19	18	19	18	16	18
1984	15	14	15	14	13	14	19	20	20	20	16	20
1985	15	14	15	14	13	14	21	20	21	20	17	20

years, fewer servers are required to achieve the 90% fill rate when treating removal, transportation, and repair separately. This is as we would expect, since, if these three phases are separate, removal and transportation can occur even if there is a queue at the repair facility. By treating the separate phases together as one service operation, as in the single-stage model, we are in essence requiring a free repair facility prior to removal, in which case the repairman would be idle until the component arrived. Thus, in the situation where removal and transportation have ample servers and queueing takes place only at repair, conservative results are obtained if we treat the three separate phases as one. If indeed the service portion is made up of separate phases, a more realistic representation of the system is obtained by using the series model, in spite of the slight inaccuracy caused by invoking the infinite source assumption when calculating probabilities. With fill-rate constraints of 90% or better which guarantee relatively little congestion, it appears that the inaccuracies caused by using the infinite source assumption in the series model are negligible as compared to those that result when using a single-stage finite source model if the service actually is multistage.

ACKNOWLEDGMENTS

We are indebted to Dr. Martin J. Fischer for bringing to the authors' attention the necessity of obtaining failure point probabilities for the fill rate constraint, and to the referee for many helpful suggestions.

REFERENCES

[1] Cooper, R. B., *Introduction to Queuing Theory*, (Macmillan, New York, 1972).

[2] Cruon, R., A. Rougerie, and C. Van de Casteele, "Etude d'un circuit de Matériels Réparables. Determination du Nombre de Bancs de Réparation et du Nombre d'Equipements de Rechange Nécessaires," Rev. Francaise Automat. Informat. Recherche Opérationnelle *3*, 87–102 (1969).

[3] Gordon, W. J., and G. F. Newell, "Closed Queuing Systems," Operations Research *15*, 254–265 (1967).

[4] Gross, D., and C. M. Harris, *Fundamentals of Queueing Theory*, (Wiley, New York, 1974).

[5] Gross, D., and H. D. Kahn, "On the Machine Repair Problem with Spares under Unequal Failure Rates," Technical Memorandum Serial TM-66451, Program in Logistics, Institute for Management Science and Engineering, The George Washington University, Washington, D.C. (1976).

[6] Lureau, F., "A Queueing Theoretic Analysis of Logistics Repair Models with Spare Units," Technical Report No. 55, Department of Operations Research, Stanford University, Stanford, California (1974).

[7] Mirasol, N. M., "The Output of an $M/G/\infty$ Queuing System is Poisson," Operations Research *11*, 282–284 (1963).

[8] Mirasol, N. M., "A Systems Approach to Logistics," Operations Research *12*, 707–724 (1964).

[9] Posner, M., and B. Bernholtz, "Two Stage Closed Queueing Systems with Time Lags," Canadian Operational Research Society Journal *5*, 82–99 (1967).

[10] Posner, M., and B. Bernholtz, "Closed Finite Queueing Networks with Time Lags," Operations Research *16*, 962–976 (1968).

[11] Swersey, R. J., "Closed Queuing Systems," Report No. ORC 67-1, Operations Research Center, College of Engineering, University of California, Berkeley (1967).

[12] Taylor, J., and R. R. P. Jackson, "An Application of the Birth-Death Process to the Provision of Spare Machines," Operational Research Quarterly *5*, 95–108 (1954).

OPTIMAL INVESTMENT, PRICING AND REPLACEMENT OF COMPUTER RESOURCES

Charles H. Kriebel, Anthony A. Atkinson, Huntley W. H. Zia

Graduate School of Industrial Administration
Carnegie-Mellon University
Pittsburgh, Pennsylvania

ABSTRACT

The problem of multiple-resource capacity planning under an infinite time horizon is analyzed using a nonlinear programming model. The analysis generalizes to the long term the short-run pricing model for computer networks developed in Kriebel and Mikhail [5]. The environment assumes heterogeneous resource capacities by age (vingate), which service a heterogeneous and relatively captive market of users with known demand functions in each time period. Total variable operating costs are given by a continuous psuedoconcave function of system load, capacity, and resource age. Optimal investment, pricing, and replacement decision rules are derived in the presence of economies of scale and exogenous technological progress. Myopic properties of the decision rules which define natural (finite) planning subhorizons are discussed.

1. INTRODUCTION

During recent years, a number of articles have appeared in the management science and electronic data processing literature concerning the allocation and control of computer system resources, cf Ref. [5]. The basic issue of resource allocation and control in general is not new, of course, and both economists and accountants have studied the question for many years, e.g., [8, 10]. Notwithstanding this fact, some contemporary authors have argued (rightly or wrongly) that computer systems are different: the particular characteristics of computer technology, the heterogeneity of computer users, cost-benefit measurement, and the diversity of the literature itself inhibits (or precludes) the application of a "conventional theory" to this environment. For example, the pricing of services provided by computer system resources is often an administrative decision in practice; when arguments are advanced for marginal cost pricing, counterarguments are raised that operating costs are fixed or that joint-cost allocation is indeterminate (or arbitrary), and so on.

In a previous article [5], the general issue of pricing computer resources was discussed at some length from both the theoretical perspective of a short-run price-capacity decision model and the pragmatics of implementing the resulting decision rules. In this sequel, we extend the nonlinear programming model to derive the optimal investment, pricing, and replacement of computer resources over an infinite time horizon in the presence of economies of scale and exogenous technological progress.

Several authors have used this programming formulation to address the issue of optimal capacity selection and to infer pricing policies in an environment of varying intertemporal demand. Kriebel and Mikhail [5] considered the acquisition, use, and pricing of a multiple-resource computer network over a (short-run) fixed planning horizon where capacities, once selected, remain constant throughout. The model proposes a dynamic pricing policy and imputes the total revenue for the output onto the individual resources.

Littlechild [6] introduced the concept of individual and distinguishable vintages of a single resource being capable of producing homogeneous output. The equipment, however, is characterized by an increasing operating cost function as the vintage becomes older. Although resources were implicitly assumed to have an infinite productive life, technological improvements in subsequent vintages imparted a finite economic life to any particular vintage whereby the machine simply fell into disuse because of its relative cost inefficiency.

Baumol [2] homogenized all vintages by assuming constant technology, but aspects of physical deterioration of capacity were incorporated in the model. While Littlechild assumed assets of the "one hoss shay" type, which provides constant capacity until expiration, Baumol explicitly allowed for systematic decline in capacity.

Thus, while neither author addressed the sale of equipment as a vehicle to adjust capacity, there were *de facto* approaches to model the expiration of capacity. In the model developed below, we provide directly for the sale of capacity and derive explicitly the optimality conditions associated with the disposal decisions. We also maintain the concept of vintages and assume that operating cost is a monotonic increasing function of age. It should be noted that the physical decline of capacity can be easily adopted in our model, but the notion is deemed less applicable in the computing environment.

Beyond incorporation of the resource replacement decisions, our analysis yields a natural (finite) planning subhorizon for the general problem, and thus provides an explicit analytical linkage to the previously decomposed short-run model [5]. In the next section, we summarize the major assumptions and rationalization for applying the model to the intended environment. The demand and production functions are then described, followed by a complete specification of the infinite horizon model. In Section 4, the formal results of the analysis are given which yield the corresponding optimal decision rules. Finally, contributions of this paper are summarized and additional extensions of the basic framework are addressed.

2. ENVIRONMENTAL AND ANALYTICAL ASSUMPTIONS

The prospect of employing a "market pricing system" for the control and allocation of computer services to users depends upon several basic assumptions. To begin, we assume that users are rational decision makers in consuming services, and that this rationality is reflected by individual demand functions for computation which are known to the individual and to the center in aggregate. For example, the user's demands can be presumed to have been derived on the basis of the solution to a utility maximization problem for each individual (or "job class group"). Moreover, we assume that the level or quality of service provided to each user is completely reflected in the price-requirements relation of the demand functions for each service; i.e., service quality is constant at the price charged to each user of the service in a given time period.*

Thus, the issue of intertemporal uncertainty is ruled out at the outset. We assume that users can and will alter their pattern of use over time in response to changes in the price of an output service. Furthermore, changes in the posted prices for each output service by time period are assumed to be administratively feasible.**

The production function of the center for each output service (job class) in each time period is presumed known and depends on the input requirements of the output for a set of resources and the capacity of the individual resources available in the period. For each resource class, different vintages (e.g., equipment on hand of different ages) are assumed to be homogeneous with respect to service capability provided. Resource capacity units are assumed to be continuous and resources are perfectly divisible for purchase and disposal decisions. The structure of the model dictates that all demands are prccessed within the period of arrival. Thus, for example, if at time t a user observes a forward price in period $t+T$ that is lower than the current price, and he can delay his consumption, rationality dictates that he will. This fact rules out potential internal economies from job-splitting and ensures one-period turnaround time for jobs.†

In formulating the problem, the center seeks to plan over an infinite time horizon with the objective of maximizing the net discounted value to the firm, which is the sum of producer and consumer surplus.††We assume that total variable operating costs are pseudoconcave as a continuous function of the load on the center,‡ that economies of scale exist in operations with respect to resource capacities, that exogenous improvements in computer technology occur over time, and that existing resources economically deteriorate with age.

3. AN INFINITE HORIZON MODEL

The basic model of interest is a nonlinear programming characterization which maximizes the net discounted value to the firm over an infinite planning horizon subject to known constraints on

*As discussed in Ref. [5], separate consideration of service quality can be accommodated within the model below; it is omitted here for analytical simplicity. See also the explicit treatment of service quality in Ref. [4]. For example, quality might be reflected in the availability of different products with different prices and different demands as a function of time, or by different opportunity costs to users incorporated into either the variable operating cost functional of the model or into individual user's ability and willingness to pay for quality.

**There is no question that uncertainty, congestion, and reliability figure prominently in the environment of computer services. We would argue that this occurs primarily on a microlevel time frame, e.g., job-to-job, hourly, within shift, or daily. As the time-period interval is extended, e.g., to months or quarters, these considerations rapidly diminish in importance through aggregation, and the certainty assumption becomes less critical. The more aggregated view is our intentional reference, and, similarly, the administrative consideration for price changes is less of an issue.

†Again, we emphasize the time period interval under consideration (see above footnotes).

††The objective of maximizing the net discounted value to the firm is the conventional assumption of welfare economics; e.g., see Ref [4] for an elaboration in this context. An alternate formulation is to consider the center as a discriminating monopolist (with limitations on the monopoly power) facing a relatively captive market of demand; this analysis was illustrated in Ref [5] and the cross relations are shown below. However, the monopoly model will result in suboptimizations for the firm. In practice, most captive centers consider demand exogenous and usually price services to balance (negotiated) budget costs on an annual basis; e.g., see Ref. [9].

‡See Ref. [5], pg. 107–108.

output (job class) demands in each time period, production functions for system output in terms of resource requirements for each service, net capacity restrictions for each resource class, and feasibility requirements on the decision variables. The decision variables of the problem are the investments in resource capacities over time, the disposal (sale) of obsolete (inefficient) resources, and the unit prices to charge users in each time period. The specific notation of the model follows.

Consider a computation center environment comprised of $s=1, 2, \ldots, S$ shared resources (servicing facilities) and $j=1, 2, \ldots, J$ outputs (job classes). In each time period $t=1, \ldots, T, \ldots$, the processing of users' requests for service consists of the production of an output which imposes requirements on the shared resources of the center. Let y_{jt} represent the demand of job j in time period t, which in turn is composed of a set of resource requirements $\{y_{jts}\}$.* Let p_{jt} represent the unit price (rate) for service-charged output j at time t, which is a function of the quantity of service demanded; i.e.,

$$p_{jt}=D_{jt}(y_{jt}), \qquad j=1, \ldots, J; \qquad t=1, 2, \ldots \tag{1}$$

We assume that the "demand curve" in (1) is downward-sloping for all outputs—both individually and collectively larger quantities of service will be demanded at lower prices, and conversely. At a given point in time, the center may have more than one vintage of each resource available for production; i.e., equipment of different age depending upon the time period in which it was originally purchased. Let the subscript v correspond to the vintage of a resource in time, for $v=1, 2, \ldots, t$. We assume in a time period that the given resource requirements of each output may be processed (serviced) by any one or several of the vintages of that resource available for production in the period. In particular, the production function for each output j and time period t is given by

$$y_{jt}=F_{jt}(y_{jt1}, \ldots, y_{jts}, \ldots, y_{jtS}), \tag{2}$$

where

$$y_{jts}=\sum_{v=1}^{t} y_{jtsv}$$

for $s=1, \ldots, S$ and y_{jtsv} is the output requirements for j, t of resource category s processed on (or by) facilities of vintage v.† For the production function in (2), we assume that the marginal product

$$\partial G_m(a_n, b_n, c_r)/\partial b_n \equiv g_m^{b}(b, \cdot) \text{ or simply } g_m^{b}.$$

of a resource for each and all outputs does not depend on vintage.‡ However, the variable operating cost of output is a function of resource vintage employed, viz, variable costs increase by a factor θ per period as a resource ages, where $0 \leq \theta \leq 1$. That is, a resource cannot become less expensive to operate as it becomes older. All resource capacity acquisitions are new when purchased; at time t, only resources of vintage t may be purchased. As a convention, we assume that

*For example, the y_{jts} variables might correspond to central processor time, input/output requests, primary memory space, etc.

†For convenience and to simplify notation, we will employ the following conventions in the remainder: all functions are assumed differentiable, and, as in (1) and (2), will be designated by capital letters; their derivatives will be written

‡That is, for given s we assume $f_{jt}^{s}(y_{jts}, \cdot)=f_{jt}^{s}(y_{jts}, \cdot)\ \forall v=1, \ldots, t$.

the physical acquisitions occur just prior to the start of a period so that the capacity is available for production in the period of purchase; symmetrically, a sale of capacity in a time period precludes its availability for production in the period of sale.

For each time period $t=1, 2, \ldots$, let

ρ = the discount or interest rate;

x_{ts} = the capacity of resource s purchased in period t (of vintage t only);

q_{tsv} = the amount of capacity of resource class s and vintage v sold in time period t;

$x_{tsv}=x_{vs}-\sum_{\tau=v}^{t} q_{\tau sv}$ = the net capacity of output service produced by resource class s of vintage v in time period t;

$y_{tsv}=\sum_{j=1}^{J} y_{jtsv}$ = the total output on vintage "equipment" v of resource class s in time period t;

$K_{ts}(x_{ts})$ = the investment costs of resource capacity for resource s at time t;

$Z_{ts}(q_{tsv})$ = the salvage value of q_{tsv} in period t;

$(1+\theta)^{t-v}C_{sv}(y_{tsv}, x_{tsv})$ = the total variable operating cost in time period t for resource class s of vintage v. The function is assumed to be pseudoconcave in y_{tsv}.*

As a general statement, we assume that economies of scale in each time period and technological progress over time are reflected in the investment cost functions $K_{ts}(\cdot)$. Concerning the salvage value functions, we assume that they are resource, vintage, and time specific. Further, the salvage value of an asset is less than (or equal to) its historical cost at time of purchase, and, for a given resource, the salvage value declines with age (vintage).

The objective function is to maximize, with respect to y_{jtsv}, x_{ts}, and q_{tsv}, the discounted value to the firm given by

$$\text{(3)}\quad \Omega=\sum_{t=1}^{\infty}\left\{(1+\rho)^{-t}\left(\sum_{j=1}^{J}\int_{0}^{y_{jt}} D_{jt}(u)\,du-\sum_{s=1}^{S}\sum_{v=1}^{t}(1+\theta)^{t-v}C_{sv}(y_{tsv}, x_{tsv})\right)\right.$$
$$\left.-(1+\rho)^{-(t-1)}\left(\sum_{s=1}^{S}K_{ts}(x_{ts})-\sum_{s=1}^{S}\sum_{v=1}^{t}Z_{ts}(q_{tsv})\right)\right\}$$

subject to the capacity constraint

$$\text{(4)}\quad \sum_{j=1}^{J} y_{jtsv}\le x_{vs}-\sum_{\tau=v}^{t} q_{\tau sv};\ s=1, \ldots, S;\ v=1, \ldots, t;\ \forall t,$$

*Note that in the sequel analysis of (3) below, we assume that the discounted total cash flows are also pseudoconcave functions. That is, let H_k be differentiable pseudoconcave functions $H_k:E^n\rightarrow E^1$, $k=1, \ldots, K$ and $H(x)=\sum_{k=1}^{K} H_k(x)$ for $H:E^n\rightarrow E^1$. Then we assume, respectively, that $\nabla H(x)'(y-x)=\sum_{k=1}^{K} H_k(y)\le\sum_{k=1}^{K} H_k(x)$ for any x and y, where $\nabla H_k(x)'=(\partial H_k(x)/\partial x_1, \ldots, \partial H_k(x)/\partial x_n)$.

the non-negativity constraints

$$y_{jtsv} \geq 0,\ x_{vs} \geq 0,\ q_{tsv} \geq 0;\ \forall j, s, v, t, \tag{5}$$

the demand function in (1), and the production function in (2). Finally, let μ_{tsv} be the dual variable associated with the capacity constraint in (4).

By assumption, (3) is pseudoconcave and (4) describes a convex set. Consequently, the following Kuhn-Tucker conditions become necessary and sufficient for an optimum solution to the preceding programming problem:*

$$D_{jt}(y_{jt})f'_{jt}(y_{jts}, \cdot) \leq (1+\theta)^{t-v}c'_{sv}(y_{jts}, \cdot) + (1+\rho)^{t}\mu_{tsv},\ y_{jtsv} \bar{>} 0, \text{ for } t \geq v, \text{ and all } j, t, s, v. \tag{6}$$

$$(1+\rho)^{-(v-1)}k'_{sv} + \sum_{\tau=v}^{\infty} (1+\rho)^{-\tau}(1+\theta)^{\tau-v}c'_{sv}(\cdot, x_{vs}) \geq \sum_{\tau=v}^{\infty} \mu_{\tau sv},\ x_{vs} \bar{>} 0, \text{ for all } s, v. \tag{7}$$

$$(1+\rho)^{-(t-1)}z'_{ts} + \sum_{\tau=t}^{\infty} (1+\rho)^{-\tau}(1+\theta)^{\tau-v}c'_{sv}(\cdot, q_{tsv}) \leq \sum_{\tau=t>v}^{\infty} \mu_{\tau sv},\ q_{tsv} \bar{>} 0, \text{ for } t > v \text{ and all } t, s, v. \tag{8}$$

$$x_{vs} - \sum_{\tau=v}^{t} q_{tsv} \geq y_{tsv},\ \mu_{\tau sv} \bar{>} 0 \text{ for all } t, s, v. \tag{9}$$

It is understood for the above that direct substitution is made for the demand function in (1) and the production function in (2), and that the complimentary slackness conditions are to hold within each pair of constraints. That is, if one holds with strict inequality, the other holds with strict equality, and conversely.†

4. OPTIMAL DECISION RULES

In interpreting these results, we first note some general conditions which pertain to operation of the facilities over time.

THEOREM 1: For every resource class $s=1, \ldots, S$

1.1 Vintages are utilized in decreasing order of efficiency, which is by age with newest capacity first.

1.2 In a period where excess capacity exists on the newest vintage of the resource, no output will be scheduled on older vintages, and the dual variables μ_{tsv} are zero for all vintages.

1.3 Every vintage purchased is capacity-constrained in at least one time period.

PROOF: Suppose that more than one vintage of a resource is available for production at time t where $t \geq v_1 > v_2 \ldots \geq 1$, and for v_1, the most efficient vintage, assume that excess capacity exists; that is, from (9),

$$\sum_{j=1}^{J} y_{jtsv_1} < x_{tsv_1} \rightarrow \mu_{tsv_1} = 0.$$

*For convenience in the discussion to follow, we assume that the constraint set of the problem has been reduced such that, at boundary values, there are no redundant binding constraints. Hence, a strict interpretation of complimentary slackness is presumed valid. In general, without the assumption for problems of this type, the strict conditions may not hold.

† Note for x_{tsv} in (8) that $c'_{sv}(\cdot, q_{tsv}) = -c'_{sv}(\cdot, x_{vs})$ by the chain rule.

By (6), this implies

$$(10)\qquad f_{jt}^{s}(y_{jts},\cdot)D_{jt}(y_{jt})=(1+\theta)^{t-v_1}c_{sv_1}^{v_1}(y_{jts},\cdot).$$

Now suppose statement 1.1 above is false and for $v_2<v_1$, $y_{jtsv_1}>0$. Substituting for v_2 into (6) and combining the result with the condition for v_1 gives

$$(11)\qquad (1+\theta)^{t-v_1}c_{sv_1}^{v_1}(y_{jts},\cdot)=(1+\theta)^{t-v_2}c_{sv_2}^{v_2}(y_{jts},\cdot)+(1+\rho)^{t}\mu_{tsv_1}.$$

Since μ_{tsv_1} is strictly nonnegative, (11) implies the contradiction

$$(12)\qquad (1+\theta)^{t-v_1}c_{sv_1}^{v_1}(y_{jts},\cdot)\geq(1+\theta)^{t-v_2}c_{sv_2}^{v_2}(y_{jts},\cdot).$$

Thus, the assumption on variable operating costs generates an economic preference ordering for vintage utilization.

Statement 1.2 is a direct consequence of 1.1 and (9), since, for $t\geq v_1>v_2, \ldots, \geq 1$,

$$(13)\qquad \sum_{v=1}^{t} y_{tsv}=y_{ts}<x_{tsv_1}\rightarrow\mu_{tsv}=0 \text{ for all } v$$

and $y_{ts}=y_{tsv_1}$.

Statement 1.3 follows from the problem description and (7), since resource capacity is a decision variable

$$x_{vs}>0\rightarrow\mu_{\tau sv}>0 \text{ for at least one } \tau=v, v+1, \ldots$$

Q.E.D.

The dual variable $\mu_{\tau sv}$ may be interpreted as a quasi-rent for the period of the given vintage of the resource, and is defined in the next theorem.

THEOREM 2: The dual variable μ_{tsvk} is equal to the present value of the difference between the actual marginal operating cost for each vintage of a resource class s and the marginal operating cost of the least efficient vintage of that class employed in production at less than capacity during the time period; i.e., for $t\geq v>v^*$,

$$(14)\qquad \mu_{tsv}=(1+\rho)^{-t}\{(1+\theta)^{t-v^*}c_{sv^*}^{v^*}(y_{jts},\cdot)-(1+\theta)^{t-v}c_{sv}^{v}(y_{jts},\cdot)\}$$

PROOF: Suppose that in period t, the total requirements of a given resource exceed vintage capacities down to but not exceeding a particular vintage v^*, where $t\geq v_1>v_2>\ \ldots\ >v^*\geq 1$. That is, let

$$(15)\qquad y_{ts}=y_{tsv^*}+\sum_{\tau=v^*+1}^{v_1} x_{ts\tau}$$

From (6) and Theorem 1, we have for all $v_i>v^*$

$$(16)\qquad f_{jt}^{s}(y_{jts},\cdot)\,D_{jt}(y_{jt})=(1+\theta)^{t-v_i}c_{sv_i}^{v_i}(y_{jts},\cdot)+(1+\rho)^{t}\mu_{tsv_i},$$

and for $v_i=v^*$

$$(17)\qquad f_{jt}^{s}(y_{jts},\cdot)D_{jt}(y_{jt})=(1+\theta)^{t-v^*}c_{sv^*}^{v^*}(y_{jtv^*},\cdot).$$

Since (16) and (17) must hold simultaneously for all vintages and each resource, direct substitution yields

$$\mu_{tsv_i}=(1+\rho)^{-t}\{(1+\theta)^{t-v^*}c_{sv^*}^{v^*}(x_{tsv^*},\cdot)-(1+\theta)^{t-v_i}c_{sv_i}^{v_i}(y_{jts},\cdot))\}$$

as stated. Q.E.D.

Regarding Theorem 2, the formal interpretation is valid only when the least efficient vintage is employed at less than capacity during the time period, as stated. Under the case where the marginal vintage is also used to capacity, the equations in (14) are underidentified and the allocation may appear arbitrary, despite the fact that the v^* vintage can always be identified for each resource class in any time period on the basis of marginal operating costs.*

DEFINITION 1: Define the quasi-marginal cost (MC^+_{tsv}) of a vintage resource (s, v) as the sum of the actual marginal operating cost at total output (MC_{tsv}) and the future worth at t of the quasi-rent, $\mu_{tsv} \geq 0$.

(18) $$\mathrm{MC}^+_{tsv} \equiv (1+\theta)^{t-v} c_{vs}^{s}(y_{jts}, \cdot) + (1+\rho)^t \mu_{tsv}$$

COROLLARY (Theorem 2): In each time period, the quasi-marginal cost of a resource class s is the same for all vintages of the given resource employed in production.

PROOF: Trivial from Theorem 2: $\mathrm{MC}^+_{tsv} = \mathrm{MC}_{tsv^*} = \mathrm{MC}^+_{ts} \; \forall k$. Q.E.D.

We now state the optimal pricing decision rule for each job class.

THEOREM 3: In each time period t,

3.1 For every job, the value of the marginal product of each resource class s must equal its quasi-marginal cost;

3.2 The optimal price for each job (p^*_{jt}) is equal to the quasi-marginal cost of the job (MC^+_{jt}); i.e., for some s,

(19) $$y_{jts} > 0 \rightarrow y_{jt} > 0 \rightarrow p^*_{jt} = \mathrm{MC}^+_{jt}$$

PROOF: The first statement follows from (6) and the preceding corollary; viz.,

(20) $$D_{jt}(y_{jt}) f_{jt}^{s}(y_{jts}, \cdot) = \mathrm{MC}^+_{ts} \forall_j.$$

It is also true that for each given j, (20) holds for every s, and, further, for the quasi-marginal cost of a resource for job j,

(21) $$\mathrm{MC}^+_{jts} = \mathrm{MC}^+_{ts}$$

Multiplying both sides of (20) by dy_{jts} and summing over all resources, we obtain

(22) $$\frac{\sum_{s=1}^{S} \mathrm{MC}^+_{ts}\, dy_{jts}}{\sum_{s=1}^{S} f^s_{jt}(y_{jts}, \cdot)\, dy_{jts}} = D_{jt}(y_{jt}) = p^*_{jt}$$

We know that, for the marginal cost of each job,

(23) $$\mathrm{MC}_{jt} = dC_t / dy_{jt}$$

and the respective total differentials can be written as

(24) $$\frac{dC_t}{dy_{jt}} = \frac{\sum_{s=1}^{S} \mathrm{MC}_{jts}\, dy_{jts}}{\sum_{s=1}^{S} f^s_{jt}(y_{jts}, \cdot)\, dy_{jts}}.$$

*In this case, one can interpret μ_{tsv^*} as approximately equal to zero; however, this informal interpretation is notwithstanding the formal conditions below. That is, we obtain exact identification for all the μ_{tsv} variables with the addition of (26) below.

By analogy from (24), we can interpret the l.h.s. of (22) as the quasi-marginal cost of the job, MC_{ji}^{+}. Computation of the optimal pricing rule can be obtained from (22) or (20) since

$$(25)\qquad p_{ji}^{*}=MC_{ji}^{+}=\frac{MC_{ti}^{+}}{f_{ji}'(y_{jti},\cdot)}\text{ for all } s. \qquad \text{Q.E.D.}$$

We note that the conditions in (20) and (24) permit the imputed allocation of costs (prices) to the resources consumed by each job.

Turning our attention to (7) and (8), we determine the optimal investment and replacement decision rules.

DEFINITION 2. The long-run marginal cost of resource capacity x_{ts} is equal to the present value of the marginal investment cost plus the sum of the discounted marginal operating costs of capacity in all periods of use.

THEOREM 4. Optimal investment and replacement decision rules. Purchase x_{vs} of resource class s in time period v, and sell (retire) $q_{tvs}=x_{vs}$ in time period t, $t>v$, provided the long-run marginal cost of resource capacity in the $t-v$ periods of use less the present value of the marginal return from salvage in period t is equal to the total quasi-rent for the vintage over the planning subhorizon.

Otherwise, do not invest.

The time frame $v\leq\tau\leq t$ constitutes a natural planning subhorizon of the infinite horizon decision problem.

$$(26)\quad x_{vs}>0\rightarrow(1+\rho)^{-(v-1)}k_{vs}'+\sum_{\tau=v}^{t-1}(1+\rho)^{-\tau}(1+\theta)^{\tau-v}c_{vs}'(\cdot,x_{vs})$$
$$-(1+\rho)^{-(t-1)}z_{ts}'=\sum_{\tau=v}^{t-1}\mu_{\tau vs}>0,\ v=1,2,\ldots,;s=1,\ldots,S$$

and

$$(27)\qquad q_{tvs}>0\rightarrow q_{tvs}=x_{vs},\ v<t\leq\infty.$$

PROOF: From (7) and (9), we have $\forall v, s$

$$(7')\qquad (1+\rho)^{-(v-1)}k_{vs}'+\sum_{\tau=v}^{\infty}(1+\rho)^{-\tau}(1+\theta)^{\tau-v}c_{vs}'(\cdot,x_{vs})\geq\sum_{\tau=1}^{\infty}\mu_{\tau vs}\geq0\rightarrow x_{vs}\geqq0.$$

The marginal cost terms in the left inequality are strictly nonnegative by assumption. For decreasing values of x_{vs}, equality is approached as the successive long-run marginal cost values decrease and the associated quasi-rent values increase. Unless equality occurs for some $x_{vs}>0$, (7) and (8) have no physical meaning. Therefore, the first relation in (7′) must become a strict equality for some v and each s, and the sum of the quasi-rents will be strictly positive by Theorem 1.

Now, over the time frame $v<t\leq\tau\leq\infty$ we can substitute from (8) into (7′) for the quasi-rents; simplifying, we obtain

$$(28)\qquad (1+\rho)^{-(t-1)}z_{ts}'\leq(1+\rho)^{-(v-1)}k_{vs}'+\sum_{\tau=v}^{t-1}(1+\rho)^{-\tau}(1+\theta)^{\tau-v}c_{vs}'(\cdot,x_{vs})-\sum_{\tau=v}^{t-1}\mu_{\tau vs}$$

which is a strict inequality $\forall v$, unless $q_{tvs}>0$ for some $t>v$. The discounting process guarantees that in the limit as $t\rightarrow\infty$, (28) must become an equality by (7′). To reach equality at some $t<\infty$ in (28), the salvage term on the *l.h.s.* must decline to a slower rate over time than the *r.h.s.* for given x_{vs}; otherwise, the vintage is never sold. This implies that the total quasi-rents for the vintage must dominate the sum in all periods of use of the discounted marginal operating costs of capacity. From Theorem 2, when a vintage has become marginal and is employed at less than capacity (or is idle), $\mu_{\tau sv}=0$, $v<\tau=t$; however, marginal operating costs of capacity are positive at all output levels. Hence, equality in (28) must be obtained at or before the period t when the quasi-rents become zero; otherwise, the inequality diverges for any $0<y_{\tau sv}<x_{sv}$, $\tau>t$, in contradiction to (7′). The equality condition is restated as (26) and

$$x_{sv}>0\rightarrow q_{tsv}>0,\ v<t<\infty.$$

The time frame $v\leq\tau\leq t$ is thus a natural planning subhorizon for x_{sv}, q_{tsv} in the original infinite horizon problem by (26). Moreover, since the discounted marginal return from salvage is a decreasing function of time, and the quasi-rents in (8) have been eliminated for all $\tau>t-1$, we have (27). That is, whenever a sale is initiated, all of that vintage is sold (or, equivalently, its economic value is "written off" at time t). Q.E.D.

From (26), we see that the quasi-rents $\mu_{\tau sv}$ serve to explicitly link the market values of the outputs and the marginal costs of input resources; similarly, one can link the costs of outputs to the values of imput resources.

In interpreting the above, we observe that the marginal operating cost of capacity (i.e., $c_{sv}'(\cdot, x_{sv})$) can be either positive or negative, depending on the operating load on a given resource. This element serves to link the role of production economies of scale to the investment and disposal decisions. That is, for a low level of output, the positive c_{sv}' induces a smaller capacity investment; conversely, if the output level is high, c_{sv}' turns negative, inducing a larger investment because of savings in the operating cost at the more efficient scale. Equivalently, when output is low, the inducement is to dispose of "excess capacity" sooner; conversely, a negative c_{sv}' acts as a deterrent for the replacement decision because the discounted marginal salvage value must match the potential future savings. The natural planning subhorizon can be interpreted in the context of a myopic decision rule [1]; in this case, it provides an explicit analytical link to the previous short-run model in Ref. [5]. That is, it provides a finite planning subhorizon within the infinite time frame such that following the optimal decision rules in each subhorizon yields the optimal policy solution for the infinite horizon problem. Thus, the infinite horizon problem can be decomposed into finite subhorizon problems whose solution is facilitated by the assessment of a finite system of equations in each case.

5. CONCLUSIONS

In summary, this model is able to handle situations which previous papers have excluded. The explicit introduction of the replacement decision allows consideration of all demand possibilities between pure contraction and expansion. Incorporation of the salvage value function provides the means to decouple the infinite horizon into natural planning horizons yielding the myopic rules previously unavailable for a vintage type of model. This analysis also generalizes to the long-term the short-run pricing model in Ref. [5].

In the interests of brevity we have not repeated the earlier discussion concerning environmental characteristics of computer services and model implementation, since the comments in Ref. [5] obviously remain appropriate in this case as well. Beyond rationalizations, there appear to be three general issues that might be raised regarding the approach taken and consequently might be of interest for continuing research.

The first issue is the question of uncertainty and its resolution in the real world. Our bias here is to focus initially on the within-period uncertainty case in a somewhat more formal way than the heuristic approach outlined in Ref. [5] and as originally suggested by Smidt [8]. The goal of this effort would be to provide a model of the stochastic elements in the environment that is consistent with the overall framework. For example, under reasonable assumptions, can a characterization be developed which links to the intertemporal model through certainty equivalents or the moments of well-behaved probability distributions?

A second general issue concerns the analytical tractability of the models, particularly when one expands the framework to incorporate the within-period elements noted above. A third general issue is the empirical determination of the model. Do fundamental problems exist, such as in the specification and estimation of the economic production function of the model, which preclude the prospects for implementation from the outset?*

Clearly, these issues are interrelated and the answers to the questions may not be easy to obtain. Nevertheless, we feel that continued research is the productive next step for removing qualifications.

BIBLIOGRAPHY

[1] Arrow, K. J., "Optimal Capital Policy, The Cost of Capital, and Myopic Decision Rules," Annals of The Institute of Statistical Mathematics *16*, 21–30, Tokyo (1964).

[2] Baumol, William J., "Optimal Depreciation Policy: Pricing the Products of Durable Assets," Bell Journal of Economics and Management Science 638–656 (Autumn 1971).

[3] Fama, E., and M. Miller, *The Theory of Finance* (Holt, Rinehart and Winston, 1972).

[4] Kriebel, C. H., J. R. McCredie, A. Raviv, P. Wolk, and H. Zia, "Modeling the Productivity of Information Systems," Technical Report No. NSF APR75–20546/76/TR2, G.S.I.A. and Computation Center, Carnegie-Mellon University, Pittsburgh, Pa. (June 1976, Revised).

[5] Kriebel, Charles H., and Osama Mikhail, "Dynamic Pricing of Resources in Computer Networks," 105–124 in M. A. Geisler (ed.), *Logistics* (*North-Holland/TIMS Studies in the Management Sciences*), *1* (1975).

[6] Littlechild, S. C., "Marginal Cost Pricing with Joint Costs," Economic Journal, *80*, 223–35. (June 1970).

[7] Moring, H., "The Peak Load Problem with Increasing Returns and Pricing Constraints," American Economic Review (Sept. 1970).

[8] Pfouts, R. W., "The Theory of Cost and Production in the Multi-Product Firm," Econometrica 650–658 (October 1961).

[9] Smidt, S., "Flexible Pricing of Computer Service," Management Science B581–B600 (June 1968).

[10] Thomas, A. L., *The Allocation Problem in Financial Accounting Theory* (American Accounting Assn., Evanston, Ill., 1969).

*This research is currently in progress; e.g., see Ref. [4].

REPLACEMENT MODELS UNDER ADDITIVE DAMAGE*

Dror Zuckerman†

Cornell University
Ithaca, New York

ABSTRACT

A production system which generates income is subject to random failure. Upon failure, the system is replaced by a new identical one and the replacement cycles are repeated indefinitely. In our breakdown model, shocks occur to the system in a Poisson stream. Each shock causes a random amount of damage, and these damages accumulate additively. The failure time depends on the accumulated damage in the system. The income from the system and the cost associated with a planned replacement depend on the accumulated damage in the system. An additional cost is incurred at each failure in service. We allow a controller to replace the system at any stopping time T before failure time. We will consider the problem of specifying a replacement rule that is optimal under the following criteria: maximum total long-run average net income per unit time, and maximum total long-run expected discounted net income. Our primary goal is to introduce conditions under which an optimal policy is a control limit policy and to investigate how the optimal policy can be obtained. Examples will be presented to illustrate computational procedures.

1. INTRODUCTION AND SUMMARY

A production system is subject to a sequence of random shocks occurring in a Poisson stream at rate λ. Each shock causes a random amount of damage, and these damages accumulate additively. The successive shock magnitudes $Y_1, Y_2, \ldots$ are positive, independent, identically distributed random variables having a known distribution function $F(y)$. Any one of the shocks might cause the system to fail, the probability of which is a function of the accumulated damage caused by all previous shocks. More explicitly, if at time t the cumulative damage is $X(t)=x$, and a shock of magnitude y occurs, then the system fails with known probability $1-r(x+y)$. The function $r(\cdot)$ is referred to as the survivorship function. It will be assumed that $r(\cdot)$ is a nonincreasing function of the cumulative damage. Upon failure, the system must be replaced by a new one having the same properties, and the replacement cycles are repeated indefinitely.

Each replacement costs $C(x)$ dollars, where x is the cumulative damage at the time of replacement, and each failure adds an additional cost. The mean rate of income, when the system is *operating* and the cumulative damage is x, will be denoted by $I(x)$. The act of replacing the system requires a known amount of time. Let τ_f be the downtime associated with a failure replacement,

*This research was supported in part by the National Science Foundation under Grant Eng-7500-570.

†Now at the Hebrew University of Jerusalem.

and let τ_p be the downtime associated with a planned replacement. The accumulated damage in the system is observable. We allow a controller to institute a planned replacement at any stopping time $T<\delta$, where δ is the failure time of the system.

We consider the problem of specifying a replacement rule that is optimal under the following criteria:

(a) Maximum total long-run average net income per unit time.

(b) Maximum total long-run expected discounted net income.

Our primary goal is to introduce conditions under which an optimal policy is a control limit policy, and to investigate how the optimal policy can be obtained. The term "control limit policy" refers to a policy in which we replace either upon failure or when the accumulated damage first exceeds a critical control level ξ^*.

Barlow and Proschan [1] discussed in considerable detail the problem of finding an optimal replacement rule when the time duration that the system is in service is the only information available to the controller. Esary, Marshall and Proschan [4] investigate the property of a breakdown model for which the instants at which damage to the system occurs are Poisson distributed in time and the magnitude of damage caused by each shock equals one. Taylor [7] derives an optimal replacement rule which maximizes the total long-run average net income per unit time for the case when $C(x)$ and $I(x)$ are constants and the cumulative damage process is a compound Poisson process.

Section 2 treats the breakdown model under the long-run average net income criterion. An example will be presented to illustrate computational procedures. Section 3 considers the same breakdown model under a discounted cost criterion.

The following will be standard notation used throughout the paper: $E_x[\cdot]=E[\cdot|X(0)=x]$, $P_x(\cdot)=P(\cdot|X(0)=x)$, and reserve E (P) without affixes for expectation (probability) conditional on $X(0)=0$. The notation $E[Y;A]$, where Y is a random variable and A is an event, refers to the expectation $E[I_AY]=E[Y|I_A=1]P(A)$, where I_A is the set characteristic function of A.

2. OPTIMAL REPLACEMENT UNDER THE LONG-RUN AVERAGE NET INCOME CRITERION

Let $\{X(t);\ 0\leq t<\delta\}$ be the stochastic process described in the previous section, interpreted as the cumulative damage process up to the failure time of some production system. We allow a controller to institute a planned replacement at any stopping time $T<\delta$. Upon failure, the system must be replaced by a new identical one, and the replacement cycles are repeated indefinitely. The mean income rate as a function of the cumulative damage in the system, $I(x)$, is bounded over the state space of the damage process and nonincreasing in x. A failure replacement, the event $\{T=\delta\}$, costs K dollars. The cost of a planned replacement as a function of the cumulative damage at replacement time, $C(x)$, is assumed to be bounded and nondecreasing in x, with $C(0)=0$ and $C(x)\leq K$ for every x contained in the state space of the damage process.

To obtain the long-term expected net income per unit time, consider the renewal process formed by repeated replacements of identical systems. Theorem 3.16 of Ross [5] implies that the long-run average net income per unit time is the expected net income over a replacement cycle divided by the expected duration between replacements. That is, the average net income associated with a stopping time T will be

$$\psi_T=\frac{E\left[\int_0^T I(X(s))ds\right]-E[C(X(T));T<\delta]-KP\{T=\delta\}}{E[T]+\tau_pP\{T<\delta\}+\tau_fP\{T=\delta\}}. \tag{2.1}$$

Let $\psi^* = \sup_T \psi_T$ be the optimal average net income rate. Let ψ_L^* and ψ_U^* be lower and upper bounds respectively for ψ^*. Also let

$$R(x) = \int r(x+y)dF(y).$$

Consider the following two cases:

CASE 1: $\tau_p \leq \tau_f$, and

$$\eta(x) = \lambda(K + \psi_L^*(\tau_f - \tau_p))[1 - R(x)] - \lambda[C(x) - \int C(x+y)r(x+y)dF(y)] - I(x) \tag{2.2}$$

is nondecreasing in x.

CASE 2: $\tau_p > \tau_f$, and

$$\phi(x) = \lambda(K + \psi_U^*(\tau_f - \tau_p))[1 - R(x)] - \lambda[C(x) - \int C(x+y)r(x+y)dF(y)] - I(x) \tag{2.3}$$

is nondecreasing in x.

In this section, we will show that in both cases above, an optimal stopping time T^* is determined by a single control value ξ^*. The optimal strategy calls for replacement upon failure or when the cumulative damage first exceeds ξ^*, whichever occurs first; i.e.,

$$T^* = \min\{\inf\{t \geq 0;\ X(t) \geq \xi^*\},\ \delta\}. \tag{2.4}$$

Furthermore, we will show that ξ^* and ψ^* can be determined by solving together

$$\xi^* = \inf\{x;\ \psi^* + \lambda(K + \psi^*(\tau_f - \tau_p))[1 - R(x)] - \lambda[C(x) \\ - \int C(x+y)r(x+y)dF(y)] - I(x) \geq 0\} \tag{2.5}$$

and

$$\psi^*\{E[T^*] + \tau_p P\{T^* < \delta\} + \tau_f P\{T^* = \delta\}\} - E\left[\int_0^{T^*} I(X(s))\,ds\right] \\ + E[C(X(T^*));\ T^* < \delta] + KP\{T^* = \delta\} = 0, \tag{2.6}$$

where T^* in (2.6) has to be represented as a function of ξ^*, as will be demonstrated in Example 1.

Before proceeding, let us consider the infinitesimal operator A of the damage process. The domain of A is the set of bounded Borel measurable functions f defined on $[0, \infty)$. Let f be in the domain of A. Using the definition of the operator $A_f(x)$ (see Dynkin [3], p. 3), we obtain

$$\begin{aligned} A_f(x) &= \lim_{t \downarrow 0} t^{-1}\{E_x[f(X(t));\ t < \delta] - f(x)\} \\ &= \lim_{t \downarrow 0} t^{-1}\{(1 - \lambda t)f(x) + \lambda t \int f(x+y)r(x+y)dF(y) + o(t) - f(x)\} \\ &= -\lambda[f(x) - \int f(x+y)r(x+y)dF(y)]. \end{aligned} \tag{2.7}$$

A key tool for us is the following formula:

$$E_x[f(X(T));\ T < \delta] - f(x) = E_x\left[\int_0^T A_f(X(s))\,ds\right], \tag{2.8}$$

valid for any f in the domain of A and any stopping time T having finite expectation (see Theorem 5.1 and its corollary in Dynkin [3]).

To begin the derivation, note that for every stopping time T

$$\psi^* \geq \frac{E\left[\int_0^T I(X(s))\,ds\right] - E[C(X(T));\ T < \delta] - KP\{T = \delta\}}{E[T] + \tau_p P\{T < \delta\} + \tau_f P\{T = \delta\}}.$$

A stopping time T maximizes the long-run average net income rate if and only if T minimizes θ_T, where

(2.9)

$$\theta_T \equiv (\psi^* - \psi_T)\{E[T] + \tau_p P\{T<\delta\} + \tau_f P\{T=\delta\}\}$$
$$= \psi^*\{E[T] + \tau_p P\{T<\delta\} + \tau_f P\{T=\delta\}\} - E\left[\int_0^T I(X(s))ds\right] + E[C(X(T)); T<\delta] + KP\{T=\delta\} \geq 0,$$

and the minimum value of θ_T is zero. By the strong Markov property we have

$$E[T] = E[\delta] - E[\delta - T; T<\delta] = W(0) - E[W(X(T)); T<\delta] \tag{2.10}$$

where

$$W(x) = E_x[\delta].$$

Using (2.10), (2.9) can be rearranged to give

$$\theta_T = \psi^* \tau_p - E\left[\int_0^T I(X(s))ds\right] + \{E[f(X(T)); T<\delta] - f(0)\}, \tag{2.11}$$

where

$$f(x) = -\psi^* W(x) + C(x) - \psi^* \tau_f + \psi^* \tau_p - K.$$

Since $W(x)$ and $C(x)$ are bounded as functions of x (Taylor [7] proved that $W(x)$ is bounded as a function of x), we may apply formula (2.8) to yield

$$\theta_T = \psi^* \tau_p - E\left[\int_0^T I(X(s))ds\right] + E\left[\int_0^T A_f(X(s))ds\right].$$

By applying a standard renewal argument the following is obtained:

$$W(x) = \frac{1}{\lambda} + \int W(x+y) r(x+y) dF(y). \tag{2.12}$$

Using equations (2.7) and (2.12), we obtain

$$A_f(x) = \psi^* - \lambda[C(x) - \int C(x+y) r(x+y) dF(y)] + \lambda(K + \psi^*(\tau_f - \tau_p))[1 - R(x)].$$

We can now write

$$\theta_T = \psi^* \tau_p + E\left[\int_0^T J(X(s))ds\right], \tag{2.13}$$

where $J(x) \equiv A_f(x) - I(x)$.

Our goal is to minimize θ_T as a function of T. First consider the case $\tau_p \leq \tau_f$. $J(x)$ can be rearranged to give

$$J(x) = \eta(x) + \lambda(\psi^* - \psi_L^*)(\tau_f - \tau_p)[1 - R(x)] + \psi^*, \tag{2.14}$$

where $\eta(x)$ is given by (2.2).

Assume that $\eta(x)$ is nondecreasing. It can then be seen that $J(x)$ is also nondecreasing in x. Now, since $\xi^* = \inf\{x; J(x) \geq 0\}$ (see (2.5)), $J(x) \geq 0$ if and only if $x \geq \xi^*$. Thus, by definition of T^* (see (2.4)), we obtain for all $t<\delta$

$$J(X(t)) < 0 \text{ if and only if } t < T^*.$$

For every stopping time T,

$$(2.15)\qquad \begin{aligned}\theta_{T^*}-\theta_T&=E\left[\int_0^{T^*} J(X(s))\,ds\right]-E\left[\int_0^{T} J(X(s))\,ds\right]\\ &=E\left[\int_T^{T^*} J(X(s))\,ds;\,T<T^*\right]-E\left[\int_{T^*}^{T} J(X(s))\,ds;\,T\geq T^*\right]\leq 0.\end{aligned}$$

Thus, T^* minimizes θ_T and (2.6) expresses the boundary condition $\theta_{T^*}=0$. This completes the proof of the optimality of T^* in Case 1. By repeating the same argument, we can easily establish the optimality of T^* in Case 2.

REMARKS:

1. The optimal critical level ξ^* is nonincreasing as a function of $(\tau_f-\tau_p)$.
2. The main result of Taylor [7] can be obtained as a special case from equations (2.4), (2.5), and (2.6).
3. If $C(x)$ is a constant and $\tau_f\geq\tau_p$, an optimal policy is a control limit policy.
4. In some cases, it is possible to express

$$E\left[\int_0^{T^*} I(X(s))\,ds\right]$$

analytically as a function of ξ^* (see (2.6)), as will be demonstrated in Example 1; otherwise, simulation methods are needed.

5. It turns out that even if $I(x)$ is nonincreasing and $C(x)$ is nondecreasing, an optimal policy is not necessarily a control limit policy. To be more specific let us consider the following example (the downtimes associated with a planned replacement and with a failure replacement are negligible):

$$r(z)=\begin{cases}1 \text{ if } z<3,\\ 0 \text{ if } z\geq 3.\end{cases}$$

$F(y)=0$ for $y<1$, $1/2$ for $1\leq y<2$ and 1 for $y\geq 2$.

$$C(x)=\begin{cases}0 & \text{for } x=0,\\ 1 & \text{for } 0<x\leq 1,\\ 1+199\,(x-1) & \text{for } 1<x<2,\\ 200 & \text{for } x\geq 2.\end{cases}$$

$K=200$, $\lambda=1$, and $I(x)=1000$.

The state space of the damage process contains the states $\{0, 1, 2\}$. The theory of optimal stopping in Markov processes (Theorem 3 of Taylor [6]) implies that the optimal policy is the best one among the following set of policies:

$\pi_{(1)}$—replace at $x=1$, or at failure point of time.

$\pi_{(2)}$—replace only at failure time.

$\pi_{(3)}$—replace at $x=2$, or at failure time.

$\pi_{(4)}$—replace at $x=1, 2$.

It can be seen that the policy $\pi_{(1)}$ is the optimal policy, but clearly it is not a control limit policy.

EXAMPLE 1: Let us consider the following case:

$$r(z)=\begin{cases}1 \text{ if } z<L,\\ 0 \text{ if } z\geq L,\end{cases}$$

$$C(x)=ax, \qquad (a>0),$$

$$I(x)=b-\frac{b}{L}x, \qquad (b>0),$$

$$F(y)=1-e^{-\mu y}, \qquad (y\geq 0),$$

$$K=2aL,$$

$$\tau_p\leq\tau_f.$$

To avoid trivialities, we assume that the optimal net income rate ψ^* is positive. In view of this, we can choose 0 as a lower bound for ψ^*. Also suppose that $L\geq 1/\mu$, which is a reasonable assumption. In order to show that an optimal policy is a control limit policy, it suffices to verify monotonicity of $\eta(x)$ (see (2.2)). In our case,

$$\eta(x)=2aL\lambda\int_{L-x}^{\infty}dF(y)-\lambda\left[ax-\int_0^{L-x}a(x+y)\,dF(y)\right]-\left(b-\frac{b}{L}x\right) \tag{2.16}$$

$$=aL\lambda e^{-\mu(L-x)}-\frac{\lambda a}{\mu}e^{-\mu(L-x)}+\frac{\lambda a}{\mu}-b+\frac{b}{L}x.$$

Since $L\geq 1/\mu$, the optimal policy is a control limit policy and (ξ^*, ψ^*) can be determined by jointly solving equations (2.5) and (2.6). Clearly, $\xi^*\leq L$. Let N be the index of shock at which failure occurs. In order to express the expected income per cycle under the optimal policy as a function of ξ^*, we will introduce the substochastic kernal

$$K^n(\xi)=P\{N>n \text{ and } Y_1+Y_2+\ldots+Y_n<\xi\} \tag{2.17}$$

$$=\begin{cases}F^{(n)}(\xi) & \text{for } \xi<L,\\ F^{(n)}(L) & \text{for } \xi\geq L,\end{cases}$$

where $F^{(K)}$ is the K-fold convolution of F. Here, F is exponentially distributed with parameter μ, and so for $\xi\leq L$

$$K^n(d\xi)=\frac{1}{(n-1)!}\mu^n\xi^{n-1}e^{-\xi\mu}d\xi. \tag{2.18}$$

Since each $n<N$ has an associated intershock mean time of λ^{-1}, we obtain

$$E\left[\int_0^{T^*}I(X(s))\,ds\right]=\lambda^{-1}\left\{I(0)+\sum_{n\geq 1}\int_0^{\xi^*}I(z)K^n(dz)\right\} \tag{2.19}$$

$$=\lambda^{-1}\left\{b+b\mu\xi^*-\frac{b\mu}{2L}(\xi^*)^2\right\}.$$

The expected time between two successive replacements under the optimal policy will be

$$E[T^*]=\lambda^{-1}\left\{1+\sum_{n\geq 1}P\{N>n, \text{ and } Y_1+Y_2+\ldots+Y_n<\xi^*\}\right\} \tag{2.20}$$

$$=\lambda^{-1}\left\{1+\sum_{n\geq 1}K^n(\xi^*)\right\}=\lambda^{-1}\{1+\mu\xi^*\}.$$

By using the memoryless property of the exponential distribution we obtain

$$P\{T^*=\delta\}=e^{-\mu(L-\xi^*)} \tag{2.21}$$

and

$$E[C(X(T^*)); T^*<\delta]=\int_0^{L-\xi^*} a(y+\xi^*)\mu e^{-\mu y}\,dy \tag{2.22}$$

$$=a\xi^*+\frac{a}{\mu}-\left(\frac{a}{\mu}+aL\right)e^{-\mu(L-\xi^*)}.$$

Thus, equations (2.5) and (2.6) reduce in this case to

$$\psi^*+aL\lambda e^{-\mu(L-\xi^*)}-\frac{\lambda a}{\mu}e^{-\mu(L-\xi^*)}+\frac{\lambda a}{\mu}-b+\frac{b}{L}\xi^*+\psi^*\lambda(\tau_f-\tau_p)e^{-\mu(L-\xi^*)}=0 \tag{2.23}$$

and

$$\frac{\psi^*}{\lambda}(1+\mu\xi^*)+\psi^*\tau_p(1-e^{-\mu(L-\xi^*)})+\psi^*\tau_f e^{-\mu(L-\xi^*)} \tag{2.24}$$

$$-\frac{1}{\lambda}\left\{b+b\mu\xi^*-\frac{b\mu}{2L}(\xi^*)^2\right\}+a\xi^*+\frac{a}{\mu}-\left(\frac{a}{\mu}-aL\right)e^{-\mu(L-\xi^*)}=0$$

The optimal strategy can be determined by jointly solving equations (2.23) and (2.24).

3. OPTIMAL REPLACEMENT IN THE DISCOUNTED CASE

We now attempt to maximize the expected total discounted net income. For a given stopping time T, the expected discounted net income from the first replacement cycle is

$$U_{T,\alpha}(1)=E\left[\int_0^T e^{-\alpha s}I(X(s))\,ds\right]-E[e^{-\alpha T}C(X(T)); T<\delta]-E[e^{-\alpha T}K; T=\delta]. \tag{3.1}$$

Generally, the expected discounted net income from the n^{th} replacement cycle is

$$U_{T,\alpha}(n)=U_{T,\alpha}(1)\{e^{-\alpha\tau_p}E[e^{-\alpha T}; T<\delta]+e^{-\alpha\tau_f}E[e^{-\alpha T}; T=\delta]\}^{n-1}. \tag{3.2}$$

Clearly, we can restrict our attention to the following set of stopping times:

$$\Upsilon=\{T; X(T)\neq 0\}.$$

For each stopping time $T\in\Upsilon$,

$$E[T]\geq\frac{1}{\lambda}>0.$$

Therefore, the expected total discounted net income associated with a stopping time $T\in\Upsilon$ will be

$$U_{T,\alpha}=E\left[\lim_{n\to\infty}\sum_{i=1}^{n}\left\{\begin{matrix}\text{discounted net income associated with the } i^{th}\\ \text{replacement cycle when a stopping time } T \text{ is employed.}\end{matrix}\right\}\right].$$

By applying the dominated convergence theorem, it follows that

$$U_{T,\alpha}=\lim_{n\to\infty}\sum_{i=1}^{n}U_{T,\alpha}(i)$$

$$(3.3)\qquad =\frac{E\left[\int_0^T e^{-\alpha s} I(X(s))\,ds\right]-E[e^{-\alpha T}C(X(T));\,T<\delta]-E[e^{-\alpha T}K;\,T=\delta]}{1-e^{-\alpha\tau_p}E[e^{-\alpha T};\,T<\delta]-e^{-\alpha\tau_f}E[e^{-\alpha T};\,T=\delta]}.$$

Let $U_\alpha^* = \sup_T U_{T,\alpha}$ be the maximal total discounted net income. Let U_α^L and U_α^U be lower and upper bounds, respectively, for U_α^*. Consider the following two cases:

CASE 1:

1. $\tau_p \leq \tau_f$.
2. $C(x)$ and $r(x)$ are continuous functions over the state space of the damage process.
3. $\zeta(x)=\lambda(K+U_\alpha^L(e^{-\alpha\tau_p}-e^{-\alpha\tau_f}))[1-R(x)]-\lambda[C(x)-\int C(x+y)r(x+y)dF(y)]-I(x)-\alpha C(x)$

is nondecreasing in x.

CASE 2:

1. $\tau_p > \tau_f$.
2. $C(x)$ and $r(x)$ are continuous functions over the state space of the damage process.
3. $\sigma(x)=\lambda(K+U_\alpha^U(e^{-\alpha\tau_p}-e^{-\alpha\tau_f}))[1-R(x)]-\lambda[C(x)-\int C(x+y)r(x+y)dF(y)]-I(x)-\alpha C(x)$

is nondecreasing in x.

In both of the above cases, an optimal stopping time T^* is determined by a single control value ξ_α^*. The optimal strategy calls for replacement upon failure or when the cumulative damage first exceeds ξ_α^*, whichever occurs first. That is,

$$T^* = \min\{\inf\{t \geq 0;\, X(t) \geq \xi_\alpha^*\},\, \delta\}.$$

Furthermore,

$$(3.4)\qquad \xi_\alpha^* = \inf\{x;\, \lambda(K+U_\alpha^*(e^{-\alpha\tau_p}-e^{-\alpha\tau_f}))[1-R(x)]$$
$$-\lambda[C(x)-\int C(x+y)r(x+y)dF(y)]-I(x)-\alpha C(x)+\alpha U_\alpha^* e^{-\alpha\tau_p} \geq 0\}.$$

The above can be proved by applying the following formula:

$$(3.5)\qquad E_x[e^{-\alpha T}f(X(T))]-f(x)=E_x\left[\int_0^T e^{-\alpha s}(A_f(X(s))-\alpha f(X(s)))\,ds\right]$$

where

$$(3.6)\qquad A_f(x)=\lim_{t\downarrow 0} t^{-1}\{E_x[f(X(t))]-f(x)\}.$$

Formula (3.5) is valid for any function f such that $f(x)$ and $A_f(x)$ are bounded and continuous (see Breiman [2], p. 376). Note that equation (3.5) reduces to equation (2.8) when $\alpha = 0$.

The proof of the above described results follows a procedure similar to that used in Section 2 and therefore is omitted.

It is interesting to examine the connection between the discounted case and the undiscounted case when α approaches zero. For a given stopping time $T \in \mathcal{T}$ we have (see (3.3))

$$\lim_{\alpha\downarrow 0} \alpha U_{T,\alpha} = \frac{E\left[\int_0^T I(X((s))\,ds\right]-E[C(X(T));\,T<\delta]-KP\{T=\delta\}}{E[T]+\tau_p P\{T<\delta\}+\tau_f P\{T=\delta\}}$$

and this last expression is exactly the total long-run average net income when T is employed.

EXAMPLE 2: To illustrate computational procedures, let us consider the following simple model:

$$r(s)=\begin{cases}1 \text{ if } 0\leq s<L,\\ 0 \text{ if } \quad s\geq L.\end{cases}$$

$$I(x)=I,$$

$$C(x)=C,$$

$$K=C+a, \quad (a>0),$$

$$\tau_p<\tau_f,$$

$$U_\alpha^L=0.$$

Assume that F is exponentially distributed with parameter μ. It can be seen that the above model satisfies the conditions in Case 1 and as a result, an optimal policy is a control limit policy. It suffices to restrict attention to stopping times of the form

$$T_\xi=\min\{\inf\{t\geq 0;\, X(t)\geq\xi\},\, \delta\},\ (\xi\leq L).$$

First note that for every $\xi\leq L$

$$P\{T_\xi=\delta\}=e^{-\mu(L-\xi)}.$$

Let

$$N(\xi)=\inf\left\{K;\ \sum_{i=1}^{K+1} Y_i\geq\xi\right\}.$$

Let E_K be the event $\{N(\xi)=K\}$, $K=0, 1, 2, \ldots,$ G be the distribution function of an exponential random variable with parameter λ, and $G^{(K)}$ be the K-fold convolution of G. Then

$$F_{T_\xi}(t)=P\{T_\xi\leq t\}=\sum_{K=0}^{\infty} P\{E_K\}G^{(K+1)}(t)=\sum_{K=0}^{\infty}\frac{e^{-\mu\xi}(\mu\xi)^K}{K!}G^{(K+1)}(t); \tag{3.7}$$

hence,

$$E[e^{-\alpha T_\xi}]=\int_0^\infty e^{-\alpha t}dF_{T_\xi}(t)=\hat{\phi}(\alpha)\sum_{K=0}^{\infty}\frac{e^{-\mu\xi}(\mu\xi)^K}{K!}\{\hat{\phi}(\alpha)\}^K \tag{3.8}$$
$$=\hat{\phi}(\alpha)\,Q(\hat{\phi}(\alpha)),$$

where $\hat{\phi}$ is the Laplace transform of G and $Q(\hat{\phi}(\alpha))$ is the generating function of a Poisson distribution with parameter $\mu\xi$ at the point $\hat{\phi}(\alpha)$. It is well known that $\hat{\phi}(\alpha)=\lambda/\lambda+\alpha$, and $Q(\beta)=e^{-\mu\xi+\mu\xi\beta}$; therefore,

$$E[e^{-\alpha T_\xi}]=\frac{\lambda}{\lambda+\alpha}e^{-\mu\xi+\mu\xi\frac{\lambda}{\lambda+\alpha}}. \tag{3.9}$$

As a result of the special structure of the survivorship function and the memoryless property of the exponential distribution, it can easily be seen that $T_\xi|T_\xi<\delta$, $T_\xi|T_\xi=\delta$ and T_ξ are identically distributed. Thus, equation (3.3) reduces in this case to simply

$$U_{T_\xi,\alpha}=\frac{\frac{I}{\alpha}-\left(\frac{I}{\alpha}+C+ae^{-\mu(L-\xi)}\right)\frac{\lambda}{\lambda+\alpha}e^{-\mu\xi+\mu\xi\frac{\lambda}{\lambda+\alpha}}}{1-(e^{-\alpha\tau_p}(1-e^{-\mu(L-\xi)})+e^{-\alpha\tau_f}e^{-\mu(L-\xi)})\frac{\lambda}{\lambda+\alpha}e^{-\mu\xi+\mu\xi\frac{\lambda}{\lambda+\alpha}}}. \tag{3.10}$$

In order to find the optimal policy, we have simply to minimize $U_{T_\xi,\alpha}$ for $0 \leq \xi \leq L$.

REMARK: In some cases, it is impossible to express $U_{T_\xi,\alpha}$ analytically as a function of ξ, and simulation methods are needed.

REFERENCES

[1] Barlow, R. E., and F. Proschan, *Mathematical Theory of Reliability*, (John Wiley & Sons, New York 1965).

[2] Breiman, L., *Probability*, (Addison Wesley, Reading, Massachusetts 1968).

[3] Dynkin, E. B., *Markov Processes 1*, (Academic Press, New York 1965).

[4] Esary, J. D., A. W. Marshall, and F. Proschan, "Shock Models and Wear Processes," Annals of Probability *1*, 627–649 (1973).

[5] Ross, S. M., *Applied Probability Models with Optimization Applications*, (Holden-Day, San Francisco 1970).

[6] Taylor, H. M., "Optimal Stopping in a Markov Process," Annals of Mathematical Statistics *39*, 1333–1344 (1968).

[7] Taylor, H. M., "Optimal Replacement Under Additive Damage and Other Failure Models," Naval Research Logistics Quarterly *22*, (1) 1–18 (1975).

THE ONE-PERIOD, N-LOCATION DISTRIBUTION PROBLEM*

Uday S. Karmarkar
University of Chicago
Chicago, Illinois

Nitin R. Patel
Indian Institute of Management
Ahmedabad, India

ABSTRACT

This paper studies the one-period, general network distribution problem with linear costs. The approach is to decompose the problem into a transportation problem that represents a stocking decision, and into decoupled newsboy problems that represent the realization of demand with the usual associated holding and shortage costs. This approach leads to a characterization of optimal policies in terms of the dual of the transportation problem. This method is not directly suitable for the solution for large problems, but the exact solution for small problems can be obtained. For the numercial solutions of large problems, the problem has been formulated as a linear program with column generation. This latter approach is quite robust in the sense that it is easily extended to incorporate capacity constraints and the multiproduct case.

THE PROBLEM

The problem can be stated as

$$\text{(1)} \qquad \min_{\{O_i\}\{x_{ij}\}} \sum_i p_i O_i + \sum_i \sum_{j \neq i} c_{ij} x_{ij} + \sum_i \phi_i(s_i), \qquad \begin{array}{l} i=1, 2, \ldots, n \\ j=1, 2, \ldots, n, \end{array}$$

where

$$\text{(1a)} \qquad \phi_i(s_i) = C_{o_i} \int_0^{s_i} (s_i - \xi) f_i(\xi)\, d\xi + C_{u_i} \int_{s_i}^{\infty} (\xi - s_i) f_i(\xi)\, d\xi,$$

subject to

$$\text{(2a)} \qquad s_i = s_i^o + O_i + \sum_j x_{ji} - \sum_j x_{ij}, \ \forall_i$$

$$\text{(2b)} \qquad s_i,\ O_i,\ x_{ij} \geq 0, \qquad \forall_{ij},$$

*This work was performed under ONR Contract N00014-67-A-0204-0076 on Multilevel Logistics Organization Models at the Massachusetts Institute of Technology.

INTRODUCTION

In this paper we consider the single product, single period, multilocation inventory problem with stochastic demands and transshipment between locations. This problem was posed and investigated by Gross [4], where exact solutions were obtained for the one location and two location cases. Gross' method of solution rapidly becomes complicated to the point of intractability as the number of locations increases, and Gross suggests that search techniques be used to obtain numerical solutions for larger problems.

In an early paper, Allen [1] discussed the problem of stock redistribution. His model is a special case of the more general situation dealt with here. Krishnan and Rao in Ref. [9] have tackled a one-period problem similar to that proposed by Gross. However, while Gross' formulation considered ordering and shipping decisions made simultaneously at the start of the period, the approach here was to determine optimal ordering decisions, given that transshipment decisions could be deferred till demand was realized. An additional simplification made in this paper was to assume that all transshipment costs are equal. This allows arbitrary partitioning of the locations into groups, with the same transportation cost still holding between any two groups.

This problem has been examined in Elmaghraby [3] and Williams [12] using a stochastic programming type of approach. We will here adopt essentially the same viewpoint and demonstrate the connection with Gross' solution. The notation used here is that of Gross to facilitate comparison of the results. Das [2] has studied a similar problem. However, his investigation was limited to a specific two-location problem.

In the following sections, a decomposition methodology is developed and applied to the two-location problem as an illustrative case. The simplicity of the approach allows several cases of special two-location problems to be investigated easily and, in each case, the complete form of the optimal policy for all initial stock conditions can be exhibited (Karmarkar [6]).

where

s_{io}=Initial inventory position at warehouse i,

O_i=Quantity ordered from central location for warehouse i,

x_{ij}=Quantity transshipped from warehouse i to warehouse j,

s_i=Total quantity stocked at warehouse i after ordering and transshipment,

p_i=Variable cost of ordering for warehouse i from central location ($/unit),

c_{ij}=Transshipment cost from warehouse i to warehouse j ($/unit),

C_{oi}=Cost of overstocking at warehouse i ($/unit),

C_{ui}=Cost of understocking at warehouse i ($/unit),

$f_i(\cdot)$=Density (p.d.f.) of demand at warehouse i.

The assumptions made in this model are as follows:

(a) The demand at each warehouse is a random variable characterized by a continuous density function.

(b) Delivery of stock is immediate.

(c) Setup costs of ordering and transshipping are negligible.

(d) Inventory cannot be disposed of or salvaged. (This assumption is easily relaxed.)

(e) Purchasing, transshipping, holding, and shortage costs are all linear. With respect to the last two, we may make the weaker assumption that the one-period costs are convex.

(f) There are no capacity restrictions on warehouses.

(g) There is no restriction on the amount of supply available from the central location. It is shown elsewhere [6] that assumptions (f) and (g) can be relaxed.

It will be assumed throughout this article that for any combination of activities that is considered, the marginal cost of supply π_i to any location i is less than the backlog cost C_{Ui} at that location; furthermore, $\pi_i \geq -C_{Oi}$. This assumption is necessary for solutions to exist, and is discussed rigorously by Williams [13] and Karmarkar [6].

It is also assumed in [4] that

$$\text{i)}\quad c_{ij}=c_{ji}$$

$$\text{ii)}\quad p_i+c_{ij}>p_j, \qquad \forall i, j, \qquad i\neq j.$$

The last restriction is the so-called "triangular restriction" which says that it is always cheaper to order directly at any location, rather than to order at another location and transship. It turns out that this assumption really leads to the most general case in terms of the number of different optimal policies involved, and that the assumption can be relaxed with no change in the approach [6].

Let us first write down the Kuhn-Tucker conditions for the problem. We note that since the problem involves minimizing a convex function over a convex set, these conditions are necessary and sufficient.

$$\text{(3a)}\qquad \text{i)}\quad \phi_i'(s_i^*)+\lambda_i^*\geq 0, \qquad \forall i$$

$$\text{(3b)}\qquad p_i-\lambda_i^*\geq 0, \qquad \forall i$$

$$\text{(3c)}\qquad c_{ij}+\lambda_i^*-\lambda_j^*\geq 0, \qquad \forall i, j, \qquad i\neq j$$

$$\text{(4a)}\qquad \text{ii)}\quad [\phi'(s_i^*)+\lambda_i^*]s_i^*=0, \qquad \forall i$$

$$\text{(4b)}\qquad [p_i-\lambda_i^*]O_i^*=0, \qquad \forall i$$

$$\text{(4c)}\qquad [c_{ij}+\lambda_i^*-\lambda_j^*]x_{ij}^*=0, \qquad \forall i, j, \qquad i\neq j$$

$$\text{(5a)}\qquad \text{iii)}\quad s_i^*-s_i{}^o-O_i^*+\sum_j x_{ij}^*-\sum_j x_{ji}^*=0, \qquad \forall i$$

$$\text{(5b)}\qquad s_i^*, O_i^*, x_{ij}^*\geq 0.$$

We now reformulate the problem by separating the objective function into a linear part (LP) and a nonlinear stochastic part (NLP). The nonlinear part can be decoupled into independent newsboy problems and the linear part forms a transshipment problem (with the triangular restriction, this is a transportation problem).

NLP:

$$\text{(6)}\qquad \operatorname*{Min}_{\{s_i\}} \gamma_i s_i+\phi_i(s_i) \qquad \forall i, i=1, 2, \ldots, n.$$
$$s_i\geq 0$$

LP:

$$\text{(7)}\qquad \operatorname*{Min}_{\{O_i\}\,\{x_{ij}\}} \sum_i p_i O_i+\sum_i \sum_{j\neq i} c_{ij}x_{ij}.$$

subject to:

(6a) $$O_i+\sum_j x_{ji}-\sum_j x_{ij}=\bar{s}_i-s_i^o, \qquad \forall i$$

$$O_i,\ x_{ij}\geq 0.$$

Now we can write the Kuhn-Tucker conditions for each of this pair of problems and show that, under certain global restrictions, they are equivalent to the optimality conditions for the original problem.

NLP:

(8a) $$\phi_i'(s_i^*)+\gamma_i\geq 0, \qquad \forall i$$

(8b) $$[\phi_i'(s_i^*)+\gamma_i]s_i^*=0, \qquad \forall i$$

(8c) $$s_i^*\geq 0, \qquad \forall i.$$

LP:

i) Primal Constraint:

(9) $$\bar{s}_i-s_i^o+O_i^*+\sum_j X_{ij}^*-\sum_j X_{ji}^*=0, \qquad \forall i$$

ii) Dual Constraints:

(10a) $$\pi_i^*\leq p_i, \qquad \forall i$$

(10b) $$\pi_j^*-\pi_i^*\leq c_{ij}, \qquad \forall i,j, \qquad i\neq j$$

iii) Complementary Slackness:

(11a) $$[p_i-\pi_i^*]O_i^*=0, \qquad \forall i$$

(11b) $$[c_{ij}+\pi_i^*-\pi_j^*]X_{ij}^*=0, \qquad \forall i,j, \qquad i\neq j$$

(11c) $$O_i^*,\ X_{ij}^*\geq 0.$$

Now if we add to these conditions the requirements that $\gamma_i=\pi_i^*$ and $\bar{s}_i=s_i^*$, we see that these conditions are identical to the Kuhn-Tucker conditions for the original problem. This immediately suggests that a naive algorithm for solving the problem might consist of alternatively solving the two subproblems; i.e., using the NLP to generate targer inventory levels $\bar{s}_i$ for the LP, and letting the LP determine shadow prices π_i that are then used in the NLP as the marginal cost of purchase π_i. It appears that such a procedure will not in general converge, but suggests computational methods that are discussed elsewhere [6, 7].

Let us draw attention to the dual of the transportation subproblem. We can write the dual as

DLP:

(12) $$\operatorname*{Max}_{\{\pi_i\}} \sum_i (\bar{s}_i-s_i^o)\pi_i$$

subject to

$$\pi_i\leq p_i, \qquad \forall i$$

$$\pi_j-\pi_i\leq c_{ij}, \qquad \forall i\neq j.$$

We see that the dual has a very simple structure, as might be expected. Interpreting π_i as the marginal cost of providing an extra unit at the i^{th} location gives the dual constraints an intuitive meaning. The first set of constraints says that the marginal price should be less than or equal to the variable cost of ordering from the central loction. The second set says that the cost of providing an extra unit at the j^{th} location should be less than or equal to the cost of providing that unit at the i^{th} location and then transshipping it to the j^{th} location at a cost c_{ij}.

To obtain some insight into the structure of the problem, we now examine the two-location problem in some detail.

THE TWO-LOCATION PROBLEM

Let us assume for the sake of simplicity and without loss of generality that the transshipment costs between the two locations are equal in either direction. We will also assume that the "triangular restriction" is operative, so that

(13)
$$c_{12}=c_{21}=c$$
$$p_1<p_2+c$$
$$p_2<p_1+c$$

Now we can visualize the problem as setting target stock levels s_1 and s_2 so as to minimize the total costs of transshipment and subsequent realization of demand. We have to set target stock levels such that the total stock in the system is not less than the starting stock, since we cannot dispose of any stock. Once the target stock levels are picked, we may think of the ordering and transshipment decisions as being made "automatically" by the LP subproblem. Thus, if we parametrize the subproblem by the target stock vector, we can rewrite the original problem in terms of just the target stock levels as the decision variables.

$$M_2\left\{\begin{array}{l}\min\limits_{s_1,s_2} w(s_1s_2)=z(s_1,s_2)+\phi_1(s_1)+\phi_2(s_2)\\ \text{subject to}\\ \qquad s_1+s_2\geq s_1^O+s_2^O\\ \qquad s_1,s_2\geq 0,\end{array}\right.$$

where

$$P_2\left\{\begin{array}{l}z(s_1,s_2)=\min\limits_{\substack{0_1,0_2\\x_{12},x_{21}}} p_10_1+p_20_2+cx_{12}+cx_{21}\\ \text{subject to}\\ \qquad 0_1-x_{12}+x_{21}=s_1-s_1^O\\ \qquad 0_2+x_{12}-x_{21}=s_2-s_2^O\\ \qquad x_{12},x_{21},0_1,0_2\geq 0.\end{array}\right.$$

Thus, we now solve the problem M_2, where the ordering and transshipment cost function $z(s_1, s_2)$ is characterized as the optimal value of the subproblem P_2. We note that we have retained the constraints in M_2 in order to ensure the existence of a feasible solution to subproblem P_2.

We shall now obtain this transshipment cost function explicitly by considering the dual of the problem P_2:

(D)
$$\begin{aligned} z(s_1, s_2) = \text{Max } (s_1 - s_1^O)\pi_1 + (s_2 - s_2^O)\pi_2 \\ -\pi_1 + \pi_2 \leq c \\ \pi_1 - \pi_2 \leq c \\ \pi_1 \qquad \leq p_1 \\ \pi_2 \leq p_2. \end{aligned}$$

The feasible region to this program is shown in Figure 1. Given a target stock vector $\underline{s} = (s_1, s_2)$, we wish to find a feasible pair (π_1, π_2) so as to maximize the objective function, that is to maximize the projection on the gradient vector of the dual objective function.

We can see by direct inspection of the dual feasible region that

(a) For $s_1 - s_1^O > 0$, $s_2 - s_2^O > 0$, (p_1, p_2) is the optimal point and

$$z(s_1, s_2) = p_1(s_1 - s_1^O) + p_2(s_2 - s_2^O).$$

(b) For $s_1 - s_1^O < 0$, $s_1 - s_1^O + s_2 - s_2^O > 0$, $(p_2 - c, p_2)$ is the optimal point and

$$z(s_1, s_2) = (p_2 - c)(s_1 - s_1^O) + p_2(s_2 - s_2^O).$$

(c) For $s_2 - s_2^O < 0$, $s_1 - s_1^O + s_2 - s_2^O > 0$, $(p_1, p_1 - c)$ is the optimal point and

$$z(s_1, s_2) = p_1(s_1 - s_1^O) + (p_1 - c)(s_2 - s_2^O).$$

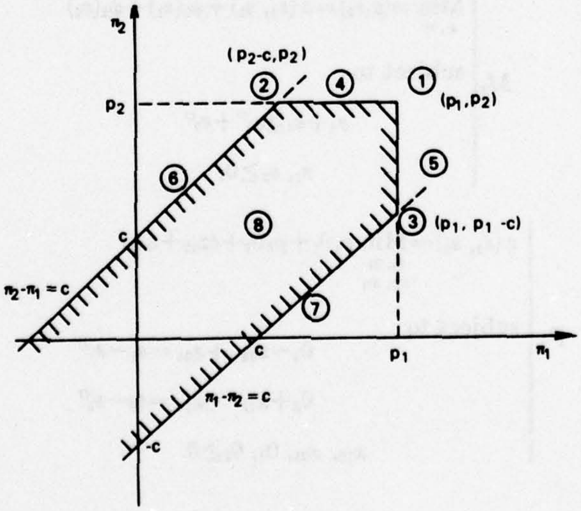


FIGURE 1. The two-location case: feasible region for the dual problem.

These represent the extreme points of the feasible region, and we get an expression for the objective function in each case, in terms of (s_1, s_2). We can also say something about the cases where the objective function is parallel to an edge of the feasible region:

(*d*) For $s_1-s_1^o=0$, $s_2-s_2^o>0$, we have $\pi_2=p_2$, $p_2-c\leq\pi_1\leq p_1$.

(*e*) For $s_1-s_1^o>0$, $s_2-s_2^o=0$, we have $\pi_1=p_1$, $p_1-c\leq\pi_2<p_2$.

(*f*) For $s_1-s_1^o+s_2-s_2^o=0$, $s_1-s_1^o<0$, $s_2-s_2^o>0$, we have $\pi_2-\pi_1=c$, $\pi_1\leq p_2-c$, $\pi_2\leq p_2$.

(*g*) For $s_1-s_1^o+s_2-s_2^o=0$, $s_1-s_1^o>0$, $s_2-s_2^o<0$, we have $\pi_1-\pi_2=c$, $\pi_1\leq p_1$, $\pi_2\leq p_1-c$.

(*h*) For $s_1-s_1^o=0$ and $s_2-s_2^o=0$, any feasible π is optimal, and finally

(*i*) For $(s_1-s_1^o+s_2-s_2^o)<0$, we have an unbounded solution to (D) corresponding to an infeasible primal.

We can represent this information about the dual variable values on the (s_1, s_2) plane. This has been shown in Figure 2. The feasible region for the original problem M is the portion of the (s_1, s_2) plane such that $s_1+s_2\geq s_1^0+s_2^0$, $s_1, s_2\geq 0$. The regions of the boundary of the dual feasible region correspond to regions in the (s_1, s_2) plane and have been marked correspondingly with the appropriate (π_1, π_2) values shown.

One method of explicitly determining these optimal policies is to visualize the problem at starting point 0 in Figure 2 with a stock position (s_1^0, s_2^0) and to move away from this point in a feasible direction to a new stock level that minimizes the total cost $w(s_1, s_2)$. The point 0 communicates with all seven regions, and it is worth moving away from it in the direction of one of these regions if the cost in that direction is decreasing. We can clearly use gradient arguments to char-

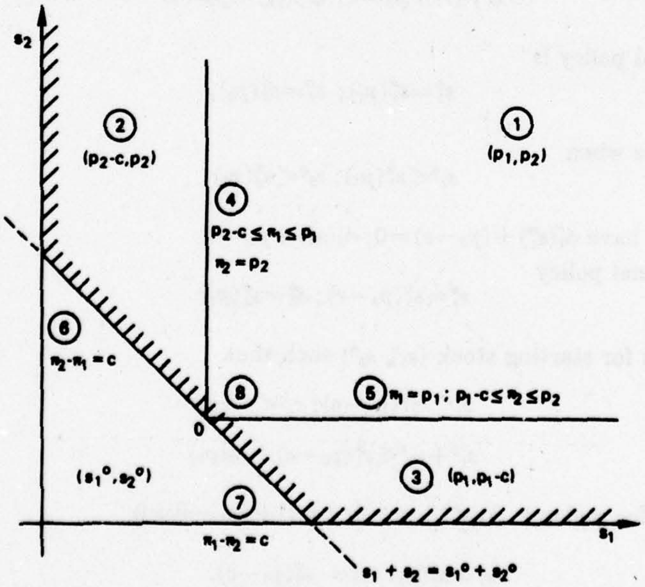


FIGURE 2. The two-location case: policy regions in the (s_1, s_2) plane.

acterize the policies, especially when we note that the derivative of the cost function in the direction of one of the axes (where it exists) is independent of the other variable; that is,

$$\nabla w(s_1, s_2) = (\phi_1'(s_1) + \pi_1, \phi_2'(s_2) + \pi_2),$$

where π_1 and π_2 depend on the region in which ∇w is evaluated. We may also note that

(a) For $s_1 > s_1{}^0$, $\pi_1 = p_1$; $s_1 < s_1{}^0$, $\pi_1 = p_2 - c$.

(b) For $s_2 > s_2{}^0$, $\pi_2 = p_2$; $s_2 < s_2{}^0$, $\pi_2 = p_1 - c$.

It is thought that this approach will prove useful in larger problems in making numerical computations. However, to determine the exact nature of the optimal policies in this case, it is simplest to directly apply the Kuhn-Tucker optimality conditions to each region, noting that the associated values of the dual variables already satisfy conditions (9) through (11*c*). So we have to consider only conditions (8*a*)$-$(8*c*) which effectively imply that either $s_i = 0$ or $\phi_i'(s_i) + \pi_i = 0$. We will assume merely for convenience that $s_i \neq 0$, although the reader may keep in mind that an optimal stock level of zero is a theoretical possibility. Also for notational ease, we define $y = s_i^*(p)$ to be the solution of

$$\phi_i(y) + p = 0.$$

We note that since $\phi_i(\cdot)$ is convex, $\phi_i'(\cdot)$ is a nondecreasing function of its argument, and it thus follows that $s_i^*(p)$ is a nondecreasing function of p. (This simply says that as the marginal cost of supplying a location increases, the optimal stock level at that location decreases.)

Now applying the $K-T$ conditions to each of the regions, we have

REGION 1: The region is defined by $s_1 > s_1{}^0$, $s_2 > s_2{}^0$.
Furthermore, we have

$$\phi'_1(s_1^*) + p_1 = 0; \ \phi'_2(s_2^*) + p_2 = 0.$$

Therefore, the optimal policy is

$$s_1^* = s_1^*(p_1); \ s_s^* = s_2^*(p_2),$$

and this policy applies when

$$s_1{}^0 < s_1^*(p_1); \ s_2{}^0 < s_2^*(p_2).$$

REGION 2: We have $\phi_1'(s_1^*) + (p_2 - c) = 0$; $\phi_2'(s_2^*) + p_2 = 0$,
which imply the optimal policy

$$s_1^* = s_1^*(p_2 - c); \ s_2^* = s_2^*(p_2).$$

This policy is optimal for starting stock $(s_1{}^0, s_2{}^0)$ such that

$$s_1{}^0 > s_1^*(p_2 - c); \ s_2{}^0 < s_2^*(p_2)$$

$$s_1{}^0 + s_2{}^0 < s_1^*(p_2 - c) + s_2^*(p_2)$$

REGION 3: As for region 2, $\phi_1'(s_1^*) + p_1 = 0$; $\phi_2'(s_2^*) + (p_1 - c) = 0$.
The optimal policy is

$$s_1^* = s_1^*(p_1); \ s_2^* = \ s_2^*(p_1 - c).$$

This policy applies if

$$s_1{}^0 < s_1^*(p_1); \ s_2{}^0 > s_2^*(p_1 - c)$$

$$s_1^0+s_2^0<s_1^*(p_1)+s_2^*(p_2-c).$$

REGION 4: We have $s_1^*=s_1^0$, and since $\phi_2'(s_2^*)+p_2=0$, we have

$$s_2^*=s_2^*(p_2).$$

Now we know that the optimal value of the dual variable π_1 must be such that $p_2-c\leq\pi_1\leq p_1$. Hence, this policy will apply when

$$\phi_1'(s_1^0)+\pi_1=0;\ s_2^0<s_2^*(p_2)$$

$$s_1^*(p_1)\leq s_1^0\leq s_1^*(p_2-c).$$

REGION 5: As in region 4, the optimal policy is

$$s_1^*=s_1^*(p_1);\ s_2^*=s_2^0.$$

This policy applies when

$$s_1^0<s_1^*(p_1);\ s_2^*(p_2)\leq s_2^0\leq s_2^*(p_1-c)$$

and the optimal value of the dual variable is given by

$$\phi_2'(s_2^0)+\pi_1=0.$$

REGION 6: The region consists of points in the feasible region such that

$$s_1<s_1^0,\ s_2>s_2^0;\ s_1+s_2=s_1^0+s_2^0.$$

For the dual variables, we have that

$$\pi_1\leq p_2-c,\ \pi_2\leq p_2;\ \pi_2-\pi_1=c,$$

and we have the conditions

$$\phi_1'(s_1^*)+\pi_1=0;\ \phi_2'(s_2^*)+\pi_2=0.$$

From the conditions on the dual variables, we have that the policy will apply when

$$s_1^0\geq s_1^*(p_2-c);\ s_1^0+s_2^0\geq s_1^*(p_2-c)+s_2^*(p_2).$$

The optimal policy is given by s_1^* and s_2^* such that

$$\phi_1'(s_1^*)-\phi_2'(s_2^*)=c$$

and

$$s_1^*+s_2^*=s_1^O+s_2^O.$$

REGION 7: The analysis here is similar to region 6. The optimal policy is given by

$$\phi_2'(s_2^*)-\phi_1'(s_1^*)=c$$

$$s_1^*+s_2^*=s_1^0+s_2^O.$$

This policy is optimal for starting stocks such that

$$s_2^O\geq s_2^*(p_1-c);\ s_1^O+s_2^O\geq s_2^*(p_1-c)+s_1^*(p_1).$$

REGION 8: Here, $s_1^*=s_1^0$, $s_2^*=s_2^0$; i.e., the optimal policy is to stay put.

The various optimal policy regions correspond to conditions on the starting stocks. We can represent these conditions on the s-plane so as to give the optimal policy as a function of the starting stocks. This is done in Figure 3, and this diagram is seen to be the same as that obtained by Gross [4]. We have thus succeeded in recovering his results.

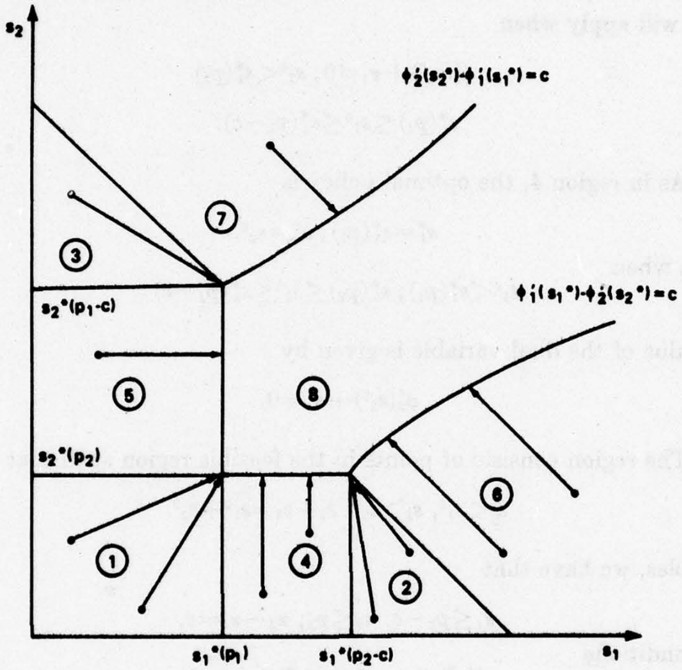


FIGURE 3. The two-location case: optimal policies for different starting stock conditions.

A NUMERICAL EXAMPLE

This example is a slight modification of the one described by Gross in Ref. [4]. The system here consists of two locations, only one of which can be supplied exogenously at a cost of \$1.50/unit. The other location can dispose of excess stock at \$1/unit. The unit shortage holding and shortage costs are \$2 and \$8, respectively, for both the locations, and costs of transshipment between locations are \$1/unit in either direction. Both locations experience uniformly distributed exogenous demands. The supply location (1) has uniform demand over 1400–2600 units, and location (2) demand is uniform over 700–1300 units. The LP dual constraints are

$$\begin{aligned} \pi_1 \quad &\leq 1.5 \\ -\pi_1+\pi_2 \quad &\leq 1.0 \\ \pi_1-\pi_2 \quad &\leq 1.0 \\ -\pi_2 \quad &\leq 1.0 \end{aligned}$$

The corresponding feasible region is sketched in Figure 4(a). For the uniform distribution, s_1^* and s_2^* are easily calculated for any given shadow prices π_1 and π_2. The nonlinear estimation problems are simple and reduce to

$$s_1^*(\pi_1)=2360-120\,\pi_1 \qquad 2180\le s_1^*\le 2600$$

$$s_2^*(\pi_2)=1180-\;60\,\pi_2 \qquad 1120\le s_1^*\le 1240.$$

For the constraints corresponding to pure transshipment policies, the equations of the lines representing optimal stock levels can be written as

$$2s_1-s_2=120$$

$$2s_1-s_2=-120$$

These equations turn out to be linear because of the uniform demand distribution assumption. The optimal policy is sketched in Figure 4(b)

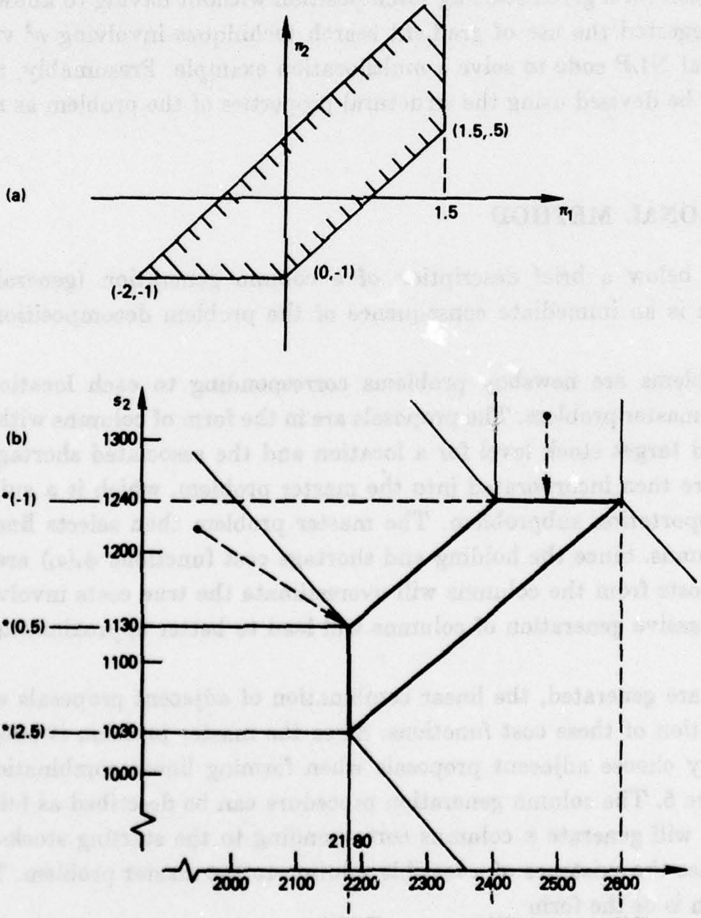


FIGURE 4. The two-location case: a numerical example.

THE FORM OF THE POLICY

We can think of the "form" of the policy described above in terms of a "static" set of points (Region 8) in which the optimal policy is to stay put. In addition, the "boundary" points of this set are also optimal stock targets for starting stocks outside the set. These starting stocks can be visualized as "cones" attached to the boundary of the "static" set by the vertex; for starting stock levels in a given cone, the optimal policy is to move to the vertex of the cone.

This kind of intuitive description of the form of the optimal policy in fact applies to a more general problem formulation allowing an arbitrary constraint set [6]. However, a complete description of the optimal policy is another matter.

While the exact method used above allowed us to specify optimal policies completely, its practicability diminishes rapidly as the number of locations increases. For one thing, geometric intuition is of little help beyond the case already discussed, and for another, the number of distinct policies to be considered becomes very large as the dimensionality of the problem increases. (There are 37 policies in the three-location case). It would clearly be useful to have a method of computing the optimal solution for a given starting stock position without having to know the whole solution. Gross [4] has suggested the use of gradient search techniques involving n^2 variables. Skeith [11] has used a general NLP code to solve a multilocation example. Presumably, more efficient search techniques could be devised using the structural properties of the problem as revealed in the analyses above.

A COMPUTATIONAL METHOD

We present below a brief description of a column generation (generalized programming) technique, which is an immediate consequence of the problem decomposition used in equations (6) and (7).

The subproblems are newsboy problems corresponding to each location, which generates proposals for the master problem. The proposals are in the form of columns with entries corresponding to a proposed target stock level for a location and the associated shortage and holding cost. These columns are then incorporated into the master problem, which is a suitably modified version of the transportation subproblem. The master problem then selects linear combinations of the proposal columns. Since the holding and shortage cost functions $\phi_i(s_i)$ are convex, the linear combination of costs from the columns will overestimate the true costs involved. This essentially ensures that successive generation of columns will lead to better approximations of the cost functions $\phi_i(\cdot)$.

As columns are generated, the linear combination of adjacent proposals will lead to an inner linear approximation of these cost functions. Since the master problem is one of minimization, it will automatically choose adjacent proposals when forming linear combinations. This procedure is shown in Figure 5. The column generation procedure can be described as follows.

Initially, we will generate n columns corresponding to the starting stocks at the n locations. This will guarantee the existence of a feasible solution to the master problem. The master problem at the t^{th} iteration is of the form

$$\min_{\{x_{ij}\},\,\{O_i\}\,\{\lambda_{ik}\}} \sum_{j\neq i} c_{ij}x_{ij}+\sum_i p_iO_i+\sum_i\sum_{k=1}^{t} c_{ik}\lambda_{ik}; \qquad i=1,\ldots,n \quad j=1,\ldots,n$$

subject to

$$\sum_j x_{ij}-\sum_j x_{ji}-O_i+\sum_{k=1}^{t} s_{ik}\lambda_{ik}=s_i^O:\pi_i$$

$$\sum_{k=1}^{t}\lambda_{ik}=1:\mu_i$$

$$O_i,\ x_{ij},\ \lambda_{ik}\geq 0.$$

We are assuming here for the sake of simplicity that one column is generated for each subproblem at each iteration. π_i and μ_i are the shadow prices corresponding to the constraints as shown. With these shadow prices, we solve n nonlinear subproblems of the form

$$\min_{s_i} \phi_i(s_i)-\pi_is_i-\mu_i; \qquad s_i\geq 0; \qquad \forall i.$$

(The negative sign on π_i is due to a sign change on the constraint.) Of course, the constant μ_i can be dropped from the objective function. These are easy to solve, and the solutions are then passed up to the master problem as columns of the form

$$\left[\begin{array}{c} c_{i,(t+1)} \\ \hdashline 0 \\ \vdots \\ s_{i,(t+1)} \\ \vdots \\ 0 \\ \hdashline 0 \\ \vdots \\ 1 \\ \vdots \\ 0 \end{array}\right]$$

The master is updated to include these columns, and then solved again. The procedure is repeated until optimality is reached or until the solution is thought to be close enough to optimality. Issues of establishing optimality, computing upper bounds on the optimal solution, and convergence of the algorithm will not be discussed here. A good exposition may be found in Lasdon [10]. Suffice it to say that the column generation algorithm will converge for any convex program.

There are several refinements possible in this procedure in the selection and incorporation of columns. For example, we may want to immediately select columns that are likely to be im-

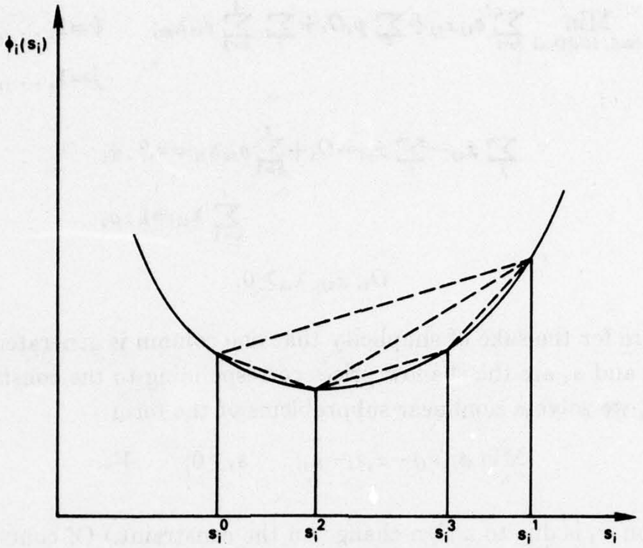


FIGURE 5. Inner linear approximation of a subproblem cost function by successive columns.

portant, corresponding to typical policies arising from extreme points of the dual. Examples of these are $s_i^*(p_i)$ and $s_i^*(p_j - c_{ji})$. Since the master problem will only use two columns at a time, it may be possible to throw out some of the columns generated after a couple of iterations. It is also straightforward to generalize the problem by adding constraints (as on capacities). These are simply introduced in the master LP as in the following example.

A 5–LOCATION EXAMPLE*

Skeith [11] describes a numerical solution, obtained by using an available NLP code, to a five-location example. His example is used here to illustrate the effectiveness of the column generation technique. Skeith's formulation differs slightly from ours in that a credit is allowed for stock carried over into the next period. However, this essentially amounts to a reduction in the cost of overstocking by the amount of credit allowed. Transshipment is allowed between the warehouses in addition to ordering from a central supply point. The centrally available stock is limited to 100 units. The demand ξ_i at the i'th location has an exponential distribution with mean $\theta_i : f_i(\xi_i) = (1/\theta_i) \exp(-\xi_i/\theta_i)$. The parameters of the problem are listed in Tables 1 and 2.

The master problem in this case is exactly as described above, with the addition of the constraint on the total stock ordered:

$$\sum_{i=1}^{5} O_i \leq 100.$$

The subproblems used to generate columns for the master problem are simplified by the exponential form of the demand distributions. Suppose that at iteration t, the master problem yields

*This example was included on the suggestion of the referee.

TABLE 1. *Five-Location Example: Cost Parameters, Initial Inventory Levels and Mean Demands*

Location i	p_i (dollars)	C_{O_i} (dollars)	C_{u_i} (dollars)	s_i^O (units)	θ_i (units)
1	2. 0	2. 0	10. 0	400	400
2	1. 0	3. 5	8. 0	450	200
3	2. 0	2. 0	12. 0	625	500
4	3. 0	0. 5	9. 0	500	300
5	2. 0	3. 0	13. 0	150	200

TABLE 2. *Five-Location Example: Transshipment Costs c_{ij} (Dollars)*

To From		Location				
		1	2	3	4	5
	1	---	1. 0	3. 0	1. 5	2. 0
	2	1. 0	---	2. 0	3. 0	2. 0
Location	3	3. 0	2. 0	---	3. 0	2. 0
	4	1. 5	3. 0	3. 0	---	1. 5
	5	2. 0	2. 0	2. 0	1. 5	---

shadow prices π_i^t. Then it can be shown that the $(t+1)$th stocking proposal for the i'th location is given by

$$s_{i,\ t+1}=-\theta_i \ln\left[(C_{O_i}-\pi_i^t)/(C_{u_i}+C_{O_i})\right].$$

The expected cost of underage and overage at the i'th location is given for the exponential case by

$$\phi_i(s_i)=C_{O_i}(s_i+\theta_i \exp(-s_i/\theta_i)-\theta_i)+C_{u_i}\theta_i \exp(-s_i/\theta_i).$$

For stock levels as calculated above, this reduces to

$$c_{i,\ t+1}=\phi_i(s_{i,\ t+1})=C_{O_i}\cdot s_{i,\ t+1}-\theta_i\cdot\pi_i^t.$$

The starting set of stocking proposals corresponded to "pure ordering" and "no change" for each location. The latter set ensured feasibility at the first iteration. The optimal solution was reached in just three iterations and the final solution was (nonzero variables only)

$$0_3=1.381, \qquad 0_5=25.065$$

$$x_{21}=154.518, \qquad x_{25}=57.565.$$

The optimal stock levels at the five locations were $s_1^*=554.518$, $s_2^*=237.917$, $s_3^*=626.381$, $s_4^*=500.00$, and $s_5^*=232.63$. The total cost of the solution was $6653. This solution differs somewhat from that published by Skeith, and it is marginally better (by $1 or so).

The master LP problems were solved on an existing interactive LP package called BIGLP, written by Linus Schrage. A small (20-line) BASIC program was written to generate columns according to the formulae above, and columns were entered interactively. Computational comparisons are not possible since Skeith's paper did not describe computational experience. However, in our case, only three LP's had to be solved, the largest of which had 11 constraints and 43 variables.

SUMMARY AND CONCLUSIONS

The approach used here is essentially simpler than that used by Gross. The key notion is that the set of points (Region 8) where the optimal policy is to stay put is obtained from the dual feasible region to the LP subproblem. That is to say, given a marginal cost of supply, the optimal stock levels can be determined. It can be easily verified that in the one location case, the solution reduces to the so-called "critical number" policy.

A rigorous treatment of the problem is given by Karmarkar [6], and the approach is extended to a general convex stochastic programming formulation. Multiperiod and infinite horizon problems, including the lost-sales case, are also discussed in Refs. [5], [6], and [8].

Special cases of two location problems can be dealt with very simply, and the optimal policies for all starting conditions exhibited. These, together with other computational methods for one-period problems, are discussed in Ref. [6].

REFERENCES

[1] Allen, S. G., "Redistribution of Total Stock over Several User Locations," Naval Research Logistics Quarterly, 5. 51–59 (1958).

[2] Das, C., "Supply and Redistribution Rules for Two-location Inventory Systems: One-Period Analysis," Management Science, 21, 765–776 (1975).

[3] Elmaghraby, S. E., "Allocation under Uncertainty When the Demand Has a Continuous D. F.," Management Science, 6, 270–294 (1960).

[4] Gross, D., "Centralized Inventory Control in Multilocation Supply Systems," Chap. 3 in H. Scarf, D. Gilford and M. Shelly (eds.): *Multistage Inventory Models and Techniques*, Stanford (University Press, Stanford, California, 1963).

[5] Karmarkar, U. S., "The Multilocation Distribution Problem," presented at the ORSA/TIMS Joint National Meeting, Chicago (1975).

[6] Karmarkar, U. S., "Multilocation Distribution Systems," unpublished Ph.D. Thesis, Massachusetts Institute of Technology (1975).

[7] Karmarkar, U. S., and N. Patel, "The One-Period, N-location Distribution Problem," O. R. Center Technical Report No. 99, Massachusetts Institute of Technology (July 1974).

[8] Karmarkar, U. S., and N. Patel, "The N-location Distribution Problem," presented at the ORSA/TIMS Joint National Meeting, Boston (1974).

[9] Krishnan, K. S., and V. R. K. Rao, "Inventory Control in N Warehouses," Journal of Industrial Engineering, *16*, 212–215 (1965).

[10] Lasdon, L. S., *Optimization Theory for Large Systems* (Macmillan, 1970).

[11] Skeith, R. W., "A Multilocation Inventory Model," Journal of Industrial Engineering, 29. 630–633 (1968).

[12] Williams, A. C., "A Stochastic Transportation Problem," Operations Research, 11, 759–770 (1963).

[13] Williams, A. C., "On Stochastic Linear Programming," SIAM Journal of Applied Mathematics, 13, 927–940 (1965).

A CONTINUOUS REVIEW INVENTORY MODEL WITH COMPOUND POISSON DEMAND PROCESS AND STOCHASTIC LEAD TIME*

Yvo M. I. Dirickx

European Institute for Advanced Studies in Management
Brussels, Belgium

Danielle Koevoets

Department of Applied Economics
Katholieke Universiteit Leuven
Leuven, Belgium

ABSTRACT

Using Markov renewal theory, we derive analytic expressions for the expected average cost associated with (s, S) policies for a continuous review inventory model with a compound Poisson demand process and stochastic lead time, under the (restrictive) assumption that only one order can be outstanding.

1. A CONTINUOUS REVIEW INVENTORY MODEL.

The demand process is of the compound Poisson type; i.e., the probability of a cumulative demand k in a time period of length t is given by

$$\sum_{l=0}^{k} e^{-\lambda t} \frac{(\lambda t)^l}{l!} v_k^{(l)},$$

where $\lambda > 0$ and $v_k^{(l)}$ is the l-fold convolution of the demand-size distribution at k; $v_k^{(1)} = v_k$, with $v_0 = 0$, $v_k > 0$ $k = 1, 2, \ldots$, denotes, of course, the probability that once a request is made on the inventory system, k items are demanded.†

The ordering process is also assumed to be stochastic. When an order is placed, the probability that the order is delivered within a time period of length t is given by the lead-time distribution

*This research is an outgrowth of the project "Sequential Decision Making" of the Department of Applied Economics, K.U.L.

†The assumption that $v_k > 0$ for all $k \geq 1$ is made for analytical convenience.

function $G(t)$, which is assumed to be continuous and Riemann integrable and to have a finite mean. (Note that $G(t)$ is assumed to be independent of the order size.)

As to the cost structure we consider

(i) an ordering cost of the form C_0+c_0k, when k items are ordered,
and per unit time:

(ii) an inventory carrying cost of c_1 per unit

(iii) a backorder cost of c_2 per unit.

Furthermore, it will be assumed that only inventory policies of the (s, S) type will be utilized. The optimization criterion will be the expected average cost criterion.

So, we want to solve the following optimization problem: Find a (s, S) policy that minimizes the expected average cost over an infinite planning horizon of the inventory problem with the demand and ordering process and the cost structure described above.*

To achieve this task, we will employ techniques of Markov renewal theory (see Cinlar [1], [2], and Schellhaas [5]) to obtain an analytic expression for the expected average cost associated with a given (s, S) policy (Sections 2 and 3) to reduce the optimization problem to a straightforward search procedure.

The proposed development will, unfortunately, necessitate the following restrictive:

ASSUMPTION: No orders can overlap.

In other words, if an order is outstanding, the next order can only be placed after the arrival of the outstanding order. It should be stressed that the tools of Markov renewal theory can be employed without this assumption; however, the difficulties arise when one attempts to derive analytical expressions for the cost rate associated with (s, S) policies.†

In order to put the present paper in perspective, we will compare the model used here with those of Gross, Harris, and Lechner (GHL) [4], Gross and Harris (GH) [3], and Schellhaas (S) [5], the latter three models being most closely related to the present one. The four models (including ours (DK)) can be classified according to the assumptions made about the demand process, the leadtime, the inventory policy, and the cost structure.

As to the demand process, (DK) is most general; in this model, it is assumed that demand can be characterized by a compound Poisson process. In (GHL) it is assumed that the demand is a compound Poisson process with the restriction, however, that the demand-size distribution assigns probability zero to demands larger than two. Finally, in (GH) and (S), a simple Poisson process is assumed to be a proper representation of the demand process.

As to the assumptions made about the leadtime, the four models split into two groups: on the one hand, (GHL) and (GH), where a "state-dependent" lead-time distribution, i.e., where the lead-time distribution is a function of the number of outstanding orders, is allowed for, and on the other hand, (S) and (DK), where the lead-time distribution is a fixed (arbitrary) distribution (with finite mean). To facilitate theoretical developments with state-dependent lead times, distribution of the exponential type are considered.

*The subsequent analysis can be adjusted to incorporate the expected discounted cost criterion.

†A manageable problem results when $(S-1, S)$ policies are employed.

With respect to the inventory policies used, (GH) and (S) use (s, S) policies. In (DK), a (s, S) policy is utilized provided no orders can overlap,* whereas a $(S-1, S)$ policy is followed in the (GH) model.

Finally, we note that (GH), (S), and (DK) are essentially equivalent in cost structure, the distinguishing feature from (GHL) being the presence of a fixed order cost. Indeed, the two types of "underage costs" considered in (GHL), (GH), and (S) can easily be incorporated in (DK) but this was not considered in (DK) for brevity of exposition.

In summary, we note that none of the models properly generalizes the others; the "no overlap" assumption prevents the present model from being the most general. We may also note that the second author is presently developing models where demand processes belong to the class of semi-Markov processes, and hence, constitute a generalization of the above models with respect to demand.

2. THE INVENTORY PROCESS AS A REGENERATIVE STOCHASTIC PROCESS.

In this section we introduce some results from Markov renewal theory in the spirit of Cinlar [1] to apply to the inventory model described above.

NOTATION

Z: integers

N: nonnegative integers

R: reals

R_+: nonnegative real numbers

Consider a probability space $(\Omega, \mathfrak{F}, P)$ and the random variables

$X_n: \Omega \to Z,$

$T_n: \Omega \to R^+,$

such that $T_0=0$, $T_n \leq T_{n+1}$ for all $n \epsilon N$ and, finally, $T_n \to \infty$ almost surely. Then

DEFINITION 1: $\{(X_n, T_n); n \epsilon N\}$ forms a Markov renewal process with state space Z if

$$P(X_{n+1}=j, T_{n+1}-T_n \leq t/X_0, \ldots, X_n; T_0, \ldots, T_n)$$
$$=P(X_{n+1}=j, T_{n+1}-T_n \leq t/X_n, T_n) \text{ for all } j \epsilon Z \text{ and } t \epsilon [0, \infty).$$

DEFINITION 2: Let $\{X(t); t \epsilon R_+\}$ be any stochastic process such that $X(t)=X_n$ for $t=T_n$. Then, $\{X(t); t \epsilon R_+\}$ is said to be a regenerative process with respect to a Markov renewal process $\{(X_n, T_n); n \epsilon N\}$ if

$$P(X(t)=j/X_0, \ldots, X_n T_0, \ldots, T_n; X(\tau) \text{ for all } \tau \leq T_n)$$
$$=P(X(t)=j/X_n, T_n) \text{ for all } t \geq T_n, n \epsilon N.$$

Now we show how the inventory problem of Section 1 can be described in terminology of Markov renewal theory. To do so, define†

(1) X_n as a random variable taking values in $J=\{S, S-1, \ldots\}$, denoting the inventory level at time period T_n; we assume $(X_0, T_0)=(k, 0)$ with $k \leq S$.

*This, in fact, rules out "state-dependent" lead time distributions.

†From now on we will assume that a particular (s, S) policy is given, in order to avoid cumbersome notation.

(2) T_n, $n \geq 1$, as the arrival time of the quantity ordered at T_{n-1} if $X_{n-1}=i$, with $i \leq s$ and as the next order point if $X_{n-1}=i$ with $s < i \leq S$.

Furthermore, in order to characterize the distribution of the length of a regeneration interval, denote by $B_j(t)$ the probability that the next regeneration point occurs in a time less than or equal to $t \epsilon R^+$ for any $j \epsilon J$.

LEMMA 1: For any $j \epsilon J$, $t \epsilon R^+$

$$B_j(t)=\begin{cases} G(t), & \text{if } j \leq s, \\ \sum_{k=-\infty}^{s} A(j, k, t), & \text{if } j > s, \end{cases}$$

with

$$A(j, k, t)=\sum_{l=0}^{j-k} e^{-\lambda t} \frac{(\lambda t)^l}{l!} v_{j-k}^{(l)}.$$

PROOF: If $j \leq s$, quantity $S-j$ is ordered; hence $B_j(t)=G(t)$. When $j > s$, observe that

$$B_j(t)=\sum_{k=-\infty}^{s} P(D_t=j-k),$$

where D_t denotes the cumulative demand in a time period of length t.

PROPOSITION 1: If X_n and T_n are defined by (1) and (2), then $\{(X_n, T_n); n \epsilon N\}$ is a Markov renewal process.

PROOF : For any $n \epsilon N$, $X_n=i$, $k \epsilon J$, consider

CASE (i) : $i > s$

$$P(X_{n+1}=k, T_{n+1}-T_n \leq t / X_0, \ldots, X_n=i; T_0, T_1, \ldots, T_n)$$

$$=\begin{cases} 0, & \text{if } k > s, \\ \int_0^t P(i-D_\tau=k / X_0, \ldots, X_n=i; T_0, T_1, \ldots, T_n) dB_i(\tau), & \text{if } k \leq s, \end{cases}$$

$$=\begin{cases} 0, & \text{if } k > s, \\ \int_0^t P(i-D_\tau=k / X_n=i, T_n) dB_i(\tau), & \text{if } k \leq s. \end{cases}$$

CASE (ii): $i \leq s$

$$P(X_{n+1}=k, T_{n+1}-T_n \leq t / X_0, \ldots, X_n=i; T_0, T_1, \ldots, T_n)$$

$$=\int_0^t P(S-D_\tau=k / X_0, \ldots, X_n=i; T_0, T_1, \ldots, T_n) dG(\tau)$$

$$=\int_0^t P(S-D_\tau=k) dG(\tau).$$

The following proposition is then immediate.

PROPOSITION 2: Let $X(t)$ be a random variable indicating the inventory level at time t. Then $\{X(t); t \geq 0\}$ is a regenerative process with respect to the Markov renewal process defined by (1) - (2).

Now that it is established that $\{(X_n, T_n); n\epsilon N\}$ is a Markov renewal process, it is, of course, well-known that $\{X_n; n\epsilon N\}$ is a Markov chain. To characterize the transition structure of this imbedded Markov chain, let

$$p_{ij}=P(X_{n+1}=j/X_n=i)$$

so that, for any pair $i, j\epsilon J$

$$(3)\qquad p_{ij}=\begin{cases}0, & \text{if } i>s,\ j>s,\\ \int_0^\infty A(i, j, \tau)\,dB_i(\tau), & \text{if } i>s\geq j,\\ \int_0^\infty A(S, j, \tau)\,dG(\tau), & \text{if } i\leq s.\end{cases}$$

In fact,

LEMMA 2: The Markov chain $\{X_n; n\epsilon N\}$ is ergodic.

PROOF: Aperiodicity and irreducibility follow from (3). To show positive recurrence, we first exhibit an explicit solution to the following system:

$$(4)\qquad \Pi_i=\sum_j p_{ji}\Pi_j, \qquad i\epsilon J,$$

$$(5)\qquad \Pi_i\geq 0, \qquad i\epsilon J,$$

$$(6)\qquad \sum_{i\epsilon J}\Pi_i=1$$

with p_{ij} as in (3).

For all $i\leq s$, we rewrite (4) as

$$(7)\qquad \Pi_i=\sum_{j=s+1}^{S} p_{ji}\Pi_j+a_i\sum_{j=-\infty}^{s}\Pi_j$$

with $a_i=p_{ji}$ for all $j\leq s$.

For $s<i\leq S$, we obtain for (4)

$$(8)\qquad \Pi_i=a_i\sum_{j=-\infty}^{s}\Pi_j.$$

Define

$$(9)\qquad \Pi_i'=\begin{cases}a_iK, & \text{if } s<i\leq S,\\ b_iK, & \text{if } i\leq s,\end{cases}$$

with

$$b_i=\sum_{j=s+1}^{S} p_{ji}a_j+a_i$$

and

$$K=\left(\sum_{i=s+1}^{S} a_i+1\right)^{-1}$$

The $\{\Pi_i'\}$ defined in (9) solve the system (5) - (8), establishing the lemma.

Lemma 2 will not only prove to be useful in subsequent theoretical developments, but also will provide us with an explicit solution (see (9)) for the stationary distribution of the Markov chain $\{X_n\}$.

In order to be able to invoke some results of Schellhaas [5], we need some definitions. Let

$$\Phi_{jk}(t, u)=P(X(t)=k/(X(o)=j, T_1=u) \text{ for } 0\leq t<u$$

and

$$\Psi_{jk}(t)=P(X(t)=k, t<T_1<\infty/X(0)=j),$$

so that

$$\Psi_{jk}(t)=\int_{t^+}^{\infty} \Phi_{jk}(t, u)dB_j(u).$$

Hence,

$$\Psi_{jk}(t)=\begin{cases} 0, & \text{for } k>j>s \text{ or } k\leq s<j, \\ \int_{t^+}^{\infty} A(j, k, t)dB_j(u), & \text{for } s<k\leq j, \\ \int_{t^+}^{\infty} A(j, k, t)dG(u) & \text{for } k\leq j\leq s. \end{cases} \tag{10}$$

The expected length of a regeneration interval with initial state j is denoted by m_j, i.e.,

$$m_j=\int_0^{\infty} t\,dB_j(t), \qquad j\epsilon J. \tag{11}$$

If $P_{jk}(t)=P(X(t)=k/X(0)=j)$, then

THEOREM 1: For any $k\epsilon J$

$$P^*_k=\lim_{t\to\infty} P_{jk}(t)=\frac{\sum_J \Pi_i \int_0^{\infty} \Psi_{ik}(t)\,dt}{\sum_J \Pi_i m_i}. \tag{12}$$

PROOF: To apply a result of Schellhaas (Ref. [5], Korrolar 1.1., p. 12), it suffices to observe that the imbedded Markov chain is ergodic (Lemma 2), that $B_j(t)$ is continuous, and that $\psi_{jk}(t)$ is Riemann-integrable.

If $V_j(t)$ denotes the expected cost associated with a particular (s, S) policy during a time interval $[0, t]$ if $X(0)=j$, our interest lies in the limiting behavior of $t^{-1}V_j(t)$. In fact, we have the following:

THEOREM 2: The limit of $t^{-1}V_j(t)$ exists for all $j\epsilon J$; in fact,

$$g_j\equiv\lim_{t\to\infty}\frac{V_j(t)}{t}=g,$$

with

$$g=g^0+g^1+g^2$$

and

$$g^0=\frac{C_0+c_0\left(S-\sum_{j\leq s} j\frac{\Pi_j}{\sum_{j\leq s}\Pi_j}\right)\sum_{j\leq s}\Pi_j}{\sum_{j=s+1}^{S}\Pi_j(m_j-m')+m'}, \tag{13}$$

$$g^1=c_1\sum_{j=0}^{S} jP^*_j, \tag{14}$$

$$g^2=c_2\sum_{j=-\infty}^{0} jP^*_j, \tag{15}$$

where the Π_j's are the solution of (4)–(6), the m_j's are defined by (11), the P_j^*'s are given in (12), and

$$m'=\int_0^\infty t\,dG(t).$$

REMARK: The fact that $g=g^0+g^1+g^2$ reflects the additive cost structure; the average cost associated with the ordering process, the inventory costs, and the backorder costs are added up to obtain the overall average costs. Note the intuitive interpretation of (13)–(15). In (13), for instance,

$$\sum_{j\leq s} j\,\frac{\Pi_j}{\sum\limits_{j\leq s}\Pi_j}$$

is the expected inventory level, given that levels below s are considered.

PROOF: The theorem is an easy consequence of a result of Schellhaas (Ref. [5], Korrolar 2.1., p. 21) which, in turn, can be established by standard renewal theoretic techniques, cf. Cinlar [2]. Schellhaas proved that, if the embedded Markov chain is ergodic, and for a cost structure which has a finite expected value on finite intervals,

$$g\equiv\lim_{t\to\infty}\frac{V_j(t)}{t}=\frac{\sum\limits_j \Pi_j\rho_j}{\sum\limits_j \Pi_j m_j}, \tag{16}$$

where ρ_j is the expected cost in a regeneration interval with initial state j.

This result will be applied to ordering, inventory carrying, and backorder costs. For the ordering costs we have

$$\rho_j{}^0=\begin{cases}0, & \text{if } j>s,\\ \sum\limits_{k=-\infty}^{S}(C_0+c_0(S-j))p_{jk}, & \text{if } j\leq s.\end{cases}$$

Hence,

$$\rho_j{}^0=\begin{cases}0, & \text{if } j>s,\\ C_0+c_0(S-j), & \text{if } j\leq s.\end{cases}$$

So,

$$g^0=\frac{\sum\limits_{j\leq s}\Pi_j(C_0+c_0(S-j))}{\sum\limits_{j=s+1}^{S}{}_j\Pi m_j+\sum\limits_{j=-\infty}^{S}\Pi_j m'}$$

and (13) is established.

Observe that for the case of the inventory carrying costs,

$$\rho_j{}^1=\begin{cases}0, & \text{if } j\leq 0,\\ \sum\limits_{k\in J}c_1\int_0^\infty \Psi_{jk}(t)\,dt, & \text{if } 0<j\leq S.\end{cases}$$

Substitution in (16) and changing the order of summation gives, then, (14). The same argument applies to the backorder costs.

Theorems 1 and 2 give us an explicit method to compute the expected average cost of a particular (s, S) policy; however, we will see in the next section how these results can be simplified from a computational point of view.

3. SOME ANALYTICAL RESULTS.

In this section, we exploit the properties of the compound Poisson process in order to obtain manageable analytical expressions for the "cost rates" (13)–(15).

In view of (9), it is clear that an expression for the stationary distribution can be obtained if the transition probabilities (see (3)) can be explicitly computed. So

LEMMA 3: For any given (s, S) policy,

$$p_{ij}=\begin{cases} 0, & \text{if } S\geq i>s,\ j>s,\\ \sum_{l=0}^{s+1-j} v_{i-j}^{(l)}2^{-(l+1)}, & \text{if } i=s+1,\ j\leq s,\\ \sum_{l=0}^{i-j} v_{i-j}^{(l)}2^{-(l+1)}-\sum_{k=1}^{i-s-1}\sum_{n=1}^{k}\frac{v_k^{(m)}}{(n-1)!}\sum_{l=0}^{i-j} v_{i-j}^{(l)}\frac{(l+n-1)!}{1!2^{l+n}}\left(1-\frac{l+n}{2n}\right) & \text{if } S\geq i\geq s+1>j,\\ \sum_{l=0}^{S-j} v_{s-j}^{(l)}\int_0^\infty e^{-\lambda t}\frac{(\lambda t)^l}{1!}\,dG(t) & \text{if } i\leq s. \end{cases}$$

PROOF: Consider the case where $S\geq i\geq s+1$ and $j\leq s$. Then

$$p_{ij}=\int_0^\infty A(i,j,t)\,d\left(1-e^{-\lambda t}-\sum_{k=1}^{j-s-1}\sum_{n=1}^{k} e^{-\lambda t}\frac{(\lambda t)^n}{n!}v_k^{(n)}\right),$$

so that

$$p_{ij}=\lambda\int_0^\infty A(i,j,t)e^{-\lambda t}dt-\lambda\sum_{k=1}^{i-s-1}\sum_{n=1}^{k} v_k^{(n)}\int_0^\infty A(i,j,t)\left(e^{\lambda t}\frac{(\lambda t)^{n-1}}{(n-1)!}-e^{-\lambda t}\frac{(\lambda t)^n}{n!}\right)dt.$$

Substituting for $A(i, j, t)$ and exploiting the properties of the gamma density gives

$$\int_0^\infty A(i,j,t)e^{-\lambda t}\frac{(\lambda t)^n}{(n)!}\,dt=\sum_{i=0}^{j-i} v_{i-j}^{(l)}\frac{(l+n)!}{\lambda l!2^{l+n+1}}.$$

Hence, the result for $S\geq i\geq s+1>j$ is established. The other cases are trivial.

We also have, in view of Lemma 1 and (11),

LEMMA 4: For any $i\epsilon J$,

$$m_i=\begin{cases} \int_0^\infty t\,dG(t), & \text{if } i\leq s,\\ \frac{1}{\lambda}, & \text{if } i=s+1,\\ \frac{1}{\lambda}\left(1+\sum_{k=1}^{j-s-1}\sum_{n=1}^{k} v_k^{(n)}\right), & \text{if } i>s+1. \end{cases}$$

Furthermore, (and this is useful to compute the values of P_k^* (see Theorem 1))

LEMMA 5: For any $i\epsilon J$

$$\int_0^\infty \Psi_{ik}(t)\,dt=\begin{cases} 0, & i>s,\ k>i \text{ or } k\leq s,\\ \sum_{k=1}^{j-s-1}\sum_{n=0}^{k} v_k^{(n)}\sum_{l=0}^{i-k} v_{i-k}^{(l)}\frac{(n+l)!}{n!l!2^{n+l+1}\lambda} & \text{if } s<k\leq i,\\ \sum_{l=0}^{i-k} v_{i-l}^{(l)}\lambda^{-1}-\sum_{l=0}^{i-k} v_{i-k}^{(l)}\int_0^\infty e^{-\lambda t}\frac{(\lambda t)^l}{l!}\,dG(t), & \text{if } k\leq i\leq s. \end{cases}$$

(For simplicity $v_k^{(0)}=1$.)
The proof of Lemma 4 and 5 is similar to that of Lemma 3.

The results can be used to evaluate (13)−(15). Unfortunately, two infinite series

$$\left(\sum_{j\le s} j\Pi_j \left(\sum_{j\le s}\Pi_j\right)^{-1} \text{and} \sum_{j=-\infty}^{0} jP_j^*\right)$$

are still to be evaluated. In practical cases, these series would have to be approximated; this is, of course, numerically speaking, not an unsurmountable task,

Further simplifications arise if a special form of the lead-time distribution is assumed. For instance, if $G(t)=1-e^{-\mu t}$, then for all $j\le s$

$$a_i=p_{ji}=\sum_{l=0}^{S-i}\frac{\mu\lambda^l}{(\lambda+\mu)^{l+1}}v_{s-i}^{(l)},$$

and setting

$$\Theta=\frac{\lambda}{\lambda+\mu},$$

$$\Pi_i=\begin{cases}\dfrac{\sum_{l=0}^{S-i}\theta^l v_{S-i}^{(l)}}{\sum_{i=s+1}^{S}\sum_{l=0}^{S-j}\theta^l v_{S-j}^{(l)}+\dfrac{\lambda+\mu}{\mu}}, & \text{if } S\geqq i\geqq s.\\[2ex] \left(\sum_{l=0}^{S-i}\dfrac{\mu}{\lambda+\mu}\theta^l v_{S-i}^{(l)}+\sum_{j=s+1}^{S}p_{ji}\sum_{l=0}^{S-j}\dfrac{\mu^l\theta^l}{(\lambda+\mu)}v_{S-j}^{(l)}\right)K & \text{if } i<s,\end{cases}$$

where K is defined as in (9). Of course, the expressions for p_{ij}, m_i, and P_i^* simplify accordingly.

The results for the case when the demand occurs according to a Poisson process can be obtained by simplifying all previous expressions to account for the fact that in this case $v_i^{(n)}=0$ whenever $i\neq n$.

If we now consider the overall expected average cost as a function of the policy used, i.e., $g(s, S)$, then it follows that $g(\cdot, S)$ is a unimodal function, and if we denote $s^*(S)$ as

$$g(s^*(S), S)=\min_{s<S} g(s, S),$$

the optimal solution is found by minimizing $g(s^*(S), S)$ over S. To achieve this task, simple search procedures can be used.

REFERENCES

[1] Cinlar, Erhan, "Markov Renewal Theory," Advances in Applied Probability *1*, 123–187 (1969).

[2] Cinlar, Erhan, "Markov Renewal Theory: A Survey," Management Science *21*, 727–752, (1975).

[3] Gross, D., and C. M. Harris, "Continuous-Review (s, S) Inventory Models with State-Dependent Leadtimes," Management Science *19*, 567–574 (1973).

[4] Gross, D., C. M. Harris, and J. A. Lechner, "Stochastic Inventory Models with Bulk Demand and State-Dependent Leadtimes," Journal of Applied Probability *8*, 521–534 (1971).

[5] Schellhaas, Helmut, "Bewertete Regenerative Prozesse mit Anwendung auf Lagerhaltungsmodelle mit Zustandsabhängigen Parametern," Preprint Nr. 136, Technische Hochschule Darmstadt, (May 1974).

A BAYESIAN APPROACH TO EVALUATING MILITARY INVESTMENTS IN PRODUCT IMPROVEMENT AND TESTING

R. T. Robinson,*

Dept. of Management
U.S. Naval War College
Newport, Rhode Island

ABSTRACT

Two issues of frequent importance in new product development are product improvement and reliability testing. A question often faced by the developer is: Should the product be distributed in its present state, or should it be improved further and/or tested before distribution? A more useful statement of the question might be: What levels of investment in further improvement and testing are economically permissible? Products for which this question is relevant may vary widely in type and intended use. This paper presents a model for determining these levels for one such product—an equipment modification procedure. The model presented makes use of present value analysis to compare cost streams and of Bayesian statistics to relate the costs to various outcomes under conditions of uncertainty. The model is applied to an actual military problem and a method is described for examining the sensitivity of the results to changes in the prior probabilities and discount rate.

INTRODUCTION

It is hardly profound to suggest that the more a decision maker knows about a process and the greater the sensitivity of the process to the resources under his control, the better and more influential his decisions are likely to be. But when the additional information and capability must be secured at a price, and when both come wrapped in a cloak of uncertainty, a more relevant question might be: "How much should the decisionmaker pay for the additional information and capability?"

The relevance of this question is perhaps most lucid in terms of a business firm developing a product for sale. The typical firm would normally undertake product improvement if the cost of the improvement program was less than the expected increase in revenue. Similarly, a firm would normally undertake a product testing program if it anticipated that the information to be gained from the testing would lead to additional marginal revenues expected to be in excess of the marginal costs of testing.

*Colonel, U.S. Army.

Thus, two distinct but related issues are involved—product improvement and product testing. The issue of product improvement, although it might well involve subjective prior estimates of expected product marketability, is rather straightforward and presents no significant computational complexities. The second issue, however—that of testing or, more generally, information gathering—is far more interesting and challenging since it involves an effort to upgrade decision quality. This effort is based not only on product improvement but also on acquiring additional information about the level of improvement that the program is expected to achieve. In other words, the first issue deals only with upgrading the product, whereas the second issue deals with upgrading the decision to market as well as product improvement.

It is also important to note that both issues involve investing money now to save money (or increase revenues) later. In other words, the typical firm that undertakes product improvement and testing is, in effect, delaying the marketing date (and the consequent revenues) in anticipation of higher revenues later from the improved product. By comparing the anticipated increases in revenue to the costs of product improvement and/or testing, the firm is able to ascertain if these efforts should be undertaken and to estimate the levels of expenditure permissible for both.

The decision procedure suggested by this discussion is relevant with respect to many decision opportunities. This paper describes the application to a decision opportunity within the military involving the development and testing of an equipment modification procedure. Three broad questions are sequentially addressed:

1. Should the modification procedure be applied in any form?
2. Should the modification procedure be applied without further development and/or testing, or should a program be undertaken to improve and/or test the procedure before field application?
3. How much is it permissible to spend on further development and testing of the modification procedure?

The decision procedure described in this paper is referred to as the Bayesian Approach because of its use of the Bayes' Probability Theorem and the Bayes' Decision Procedure (see, e.g., Raiffa [3]). Bayes' Theorem is used to develop posterior probabilities from subjectively determined prior probabilities and subsequent testing information; expected value statistics based on these probabilities are used to make expected cost comparisons among the competing alternatives. The paper concludes by presenting the process in decision tree format and by examining the sensitivity of the results to changes in certain model parameters.

THE PROBLEM EXAMINED

During the late 1960's, the U.S. military in South Vietnam began experiencing a high failure rate on a certain subassembly of one of its high-density electric power generators. Because the high failure rate had become quite costly in terms of replacement costs and generator downtime, it became necessary for the responsible logistical command to analyze the problem and consider various ways in which the situation could be corrected. The procedure described in this paper approximates the analysis that was conducted to select an appropriate corrective alternative.

The problem is presented in detail to permit the solution procedure to be developed in a rigorous manner. The reader should recognize, however, that both the problem and the procedure are very general in character and that the procedure is applicable to a wide variety of problems related to information gathering and sequential decision making under conditions of uncertainty. It is hoped that the reader will keep this generality in mind while reviewing the specific application described.

THE PRIOR PROBABILITIES OF FAILURE RATE REDUCTION

The initial stages of the decision-making process followed a rather typical pattern. For example, the engineers reviewed the technical aspects of the problem and found that a modification procedure could be applied to the generators that would result in reductions in the failure rate. Although the modification procedure could be applied without further development, the procedure was somewhat makeshift in its initial state, and the developing engineers felt that the likelihood of achieving significant failure rate reduction was relatively low. When asked to quantify their feelings, the engineers responded that they could not be too precise about the level of failure rate reduction that would be achieved, but could provide estimates of the reduction probabilities at four aggregate levels. The probability estimates provided by the engineers are listed in Table 1.

The developing engineers further indicated that if they were permitted to improve the modification procedure before it was applied, the probabilities associated with the four levels of failure rate reduction could be substantially increased. Specifically, the engineers pointed out that if they were given six more months in which to improve the procedure, the probabilities associated with the four states of failure rate reduction could be increased to the levels indicated in Table 2.

TABLE 1. *Prior Probabilities of Failure Rate Reduction Without Further Modification Development*

STATE (S_i)	LEVEL OF FAILURE RATE REDUCTION (X)	PROBABILITY ESTIMATE *OF ACHIEVING* REDUCTION LEVEL $P(S_i)$
S_1	60%	0.10
S_2	40%	0.20
S_3	20%	0.50
S_4	0%	0.20

TABLE 2. *Prior Probabilities of Failure Rate Reduction With Further Modification Development*

STATE (S_i)	LEVEL OF FAILURE RATE REDUCTION (X)	PROBABILITY ESTIMATE OF ACHIEVING REDUCTION LEVEL $P(S_i)$
S_1	60%	0.25
S_2	40%	0.50
S_3	20%	0.20
S_4	0%	0.05

SYSTEM COSTS.

The generator failures were of such a nature that personnel safety was not significantly endangered. Consequently, the analysis was based purely on cost—or, more correctly, "expected" cost, in order to recognize the future nature of such costs and the uncertainty associated with each level of failure rate reduction.

In establishing the relevant costs, it was assumed that the military intended to replace all of the generators of this type with a newer, more efficient model 3½ years from the decision date. For this reason, a 3½-year cost stream was used in determining the total cost of system repair. Furthermore, it was estimated that the cost of modifying all of the generators currently in the inventory would be $10 million for both the unimproved and the improved versions of the modification procedure. Additionally, the costs of modification improvement and testing, if undertaken either collectively or separately, would also be incurred during the first six months following the decision date; however, these costs were treated as model outputs and, hence, as unknowns in the initial cost estimates.

For ease of computation, it was assumed that if the modification was applied without further development, the modification cost of $10 million would be incurred immediately and, thus, the modified system repair costs would be incurred over the full 3½ years remaining in the generator's life cycle. Conversely, if the modification was improved before application, it was assumed that it would be applied at the end of six months from the decision date. Thus, in the latter case, present system repair costs would be incurred for the first six months and modified system repair costs would be incurred for the final three years of the generator life cycle. Finally, if the modification procedure was tested without improvement before the application decision, it was estimated that such testing would require two months and that the modification, if applied, would be applied at the end of two months from the decision date. The latter alternative and the relevant costs are discussed more fully in a later section of the paper.

With these cost data, the various cost streams were then compared using a present value analysis. A discount rate of 10 percent was used since the problem addressed involved an existing program considered to be covered by the provisions of Office of Management and Budget Circular A-94 [4] which was interpreted as requiring an annual rate of 10 percent. The equation used to compute the present value (PV) of the cost streams was

$$PV = C_0 + C_1(1+i)^{-6} + C_2(1+i)^{-18} + C_3(1+i)^{-30} + C_4(1+i)^{-42},$$

where C_i = cost associated with period i (C_0 represents initial costs),
and i = monthly discount rate.*

*The monthly discount rate, i, was selected such that the present value of a cost stream discounted monthly at $i\%$ is the same as the present value of the cost stream discounted annually at $r\%$. That is, $i = (1+r)^{1/12} - 1.0$. Thus, for $r = 10\%$, $i = 0.007974$.

The relevant cost statistics are listed in Table 3.

TABLE 3. *System Costs for All Possible Improvement Levels* ($ MILLION)

ACT	STATE (S_i)	INITIAL COST	PERIODIC REPAIR COST‡ 1	2	3	4	PRESENT VALUE
Modify without further development	S_1	10	4	8	8	8	32.78
	S_2	10	6	12	12	12	44.17
	S_3	10	8	16	16	16	55.57
	S_4	10	10	20	20	20	66.96
Modify with further development	S_1	0*	20†	8	8	8	38.04
	S_2	0*	20†	12	12	12	47.52
	S_3	0*	20†	16	16	16	57.01
	S_4	0*	20†	20	20	20	66.49
No Modify	S_4	0	10	20	20	20	56.96

*Cost of modification development and/or testing not included. The permissible expenditures for these efforts are model outputs.

†Includes six-month repair cost without modification plus $10 million for cost of modification application.

‡First period is six months in length; others are 12 months.

MODIFY OF NOT MODIFY?

The first question examined in the analysis was whether or not to undertake modification at all. The decision was based on a comparison of expected system costs over the 3½-year period with and without system modification. The expected system cost without modifying the generators was $56.96 million, which may be read directly from the last row in Table 3. The expected system cost with modification but without further development of the modification procedure was computed using the equation

$$E(C^*(=PV_1P(S_1)+PV_2P(S_2)+PV_3P(S_3)+PV_4P(S_4),$$

where C^* = system cost with modification but without further development of the modification procedure,

$P(S_i)$ = the prior probability associated with state i without further development (i=1, 2, 3, 4) (Table 1),

and PV_i = the present value of the cost stream associated with state i without further development (i=1, 2, 3, 4) (Table 3).

Thus,

$$\begin{aligned} E(C^*) &= (0.10)\ (32.78)+(0.20)\ (44.17) \\ &\quad +(0.50)\ (55.57)+(0.20)\ (66.96) \\ &= 3.28+8.83+27.79+13.39 \\ &= \$53.29 \text{ million.} \end{aligned}$$

Since the expected system cost without modification was \$56.96 million, a savings of \$3.67million (i.e., \$56.96−\$53.29) was potentially realizable by applying the modification, even without further development and testing. Thus, based on the prior probabilities given, the decision was properly to apply the modification.

FURTHER DEVELOPMENT OR NO FURTHER DEVELOPMENT?

The next decision opportunity considered was whether to apply the modification without further development or to undertake a program to improve the modification procedure before field application. This decision was made by comparing expected system costs if the modification was applied without improvement and testing, $E(C^*)$, to expected system costs if the modification was improved but not tested before application, $E(C^{**})$. The latter expected cost was computed using the equation

$$E(C^{**})=PV_1P(S_1)+PV_2P(S_2)+PV_3P(S_3)+PV_4P(S_4),$$

where C^{**} = system cost with modification and with further development of the modification procedure,

$P(S_i)$ = the prior probability associated with state i with further development (i=1, 2, 3, 4) (Table 2),

and PV_i = the present value of the cost stream associated with state i with further development (i=1, 2, 3, 4) (Table 3).

Thus,

$$\begin{aligned}E(C^{**}) &= (0.25)\ (38.04)+(0.50)\ (47.52)\\ &\quad +(0.20)\ (57.01)+(0.05)\ (66.49)\\ &= 9.51+23.76+11.40+3.33\\ &= \$48.00 \text{ million.}\end{aligned}$$

Since the expected system cost using an unimproved modification procedure was \$53.29 million, it was clearly economical to undertake the improvement program, as the potential existed for additional savings in the amount of \$5.29 million (i.e., \$53.29−\$48.00).

TESTING OR NO TESTING?

It would have been quite possible, of course, to undertake an improvement program without first subjecting the subsystem to failure testing with the improved modification applied. Accordingly, the third major question considered was: How much money should be devoted to failure (or reliability) testing in conjunction with the improvement program before making the application decision?

Clearly, the most that could be acquired from such testing was perfect information concerning the level of modification development achievable. From Table 3 it may be deduced that if perfect information had been available after development and before the application decision, the decisionmaker would have chosen to modify the generators only if the level of failure rate reduction was either state S_1 (60%) or state S_2 (40%). If the level of failure rate reduction was either state S_3 (20%) or state S_4 (0%), the decisionmaker would have chosen not to apply the modification, since with the latter two states the system cost with modification would have exceeded the system cost without modification.

The value of perfect information, then, in this problem, may be viewed as the expected savings that would have accrued to the decisionmaker by choosing not to modify if the true state was either S_3 or S_4. In terms of cost, the value of perfect information may be expressed as the difference between the expected cost if the true states were known with certainty and the expected cost when the only information about the states was the set of prior probabilities set forth in Table 2. The latter expected cost, of course, was $E(C^{**})$, or $48.00 million, as previously computed. The expected system cost corresponding to perfect information, denoted $E(C^{**}|PI)$, was computed as follows:

$$\begin{aligned} E(C^{**}|PI) &= (0.25)(38.04)+(0.50)(47.52)+ \\ &\quad (0.20)(56.96)+(0.05)(56.96) \\ &= 9.51+23.76+11.39+2.85 \\ &= \$47.51 \text{ million.} \end{aligned}$$

Therefore, the expected value of perfect information, denoted $E(PI)$, was

$$\begin{aligned} E(PI) &= E(C^{**})-E(C^{**}|PI) \\ &= 48.00-47.51 \\ &= \$0.49 \text{ million.} \end{aligned}$$

This, then, was the greatest amount that could be spent on failure testing, even if the testing could have produced perfect information. In reality, of course, no testing procedure yields perfect information. Therefore, the maximum level of expenditure allowable for testing was obviously somewhat lower than this upper bond, and, in fact, depended on the reliability of the test procedure which will be discussed in the paragraphs to follow.

THE TEST PROCEDURE.

The failure test to which the subsystem (with the improved modification applied) was to be subjected involved a testing procedure that had been used many times in the past and on which an acceptable set of reliability data had been accumulated. Experience with the test had indicated that the test results could be divided into the four groups listed in Table 4.*

TABLE 4. *Relationship Between Test Outcome and Level of Failure Rate Reduction*

TEST OUTCOME (T_j)	REDUCTION IN FAILURE RATE
T_1	60%(+)
T_2	40–60%
T_3	20–40%
T_4	0–20%

*The analysis considered the test procedure at a single fixed level. That is, the test could only be conducted or not conducted; it could not be conducted at varying levels. An interesting extension to the model would be to relax this assumption and permit alternative testing levels.

It had also been found, by comparing past test outcomes to the subsequent levels of product improvement, that once the levels of product improvement were known, the probabilities associated with the previous test results were distributed over the full range of possible test outcomes and could be estimated from the historical data. That is, if we define

$P(T_j|S_i)$=probability that the previous test result was T_j if the true state (level of improvement) turned out to be S_i,

the resulting conditional probabilities (test outcome likelihoods) for the four states and four test outcomes were derivable from past test usage. A complete listing of the likelihoods for the test is set forth in Table 5.

TABLE 5. Test Outcome Likelihoods

$P(T_j|S_i)$

STATE	TEST OUTCOMES			
	T_1	T_2	T_3	T_4
S_1	0. 900	0. 050	0. 025	0. 025
S_2	0. 050	0. 875	0. 050	0. 025
S_3	0. 025	0. 050	0. 875	0. 050
S_4	0. 025	0. 025	0. 050	0. 900

Next, the probabilities listed in Tables 2 and 5 were used to compute the joint probabilities that each combination of test outcome and failure rate reduction level would occur. These probabilities were developed from the following expression for the test outcome likelihoods:

$$P(T_j|S_i)=\frac{P(\text{Both } S_i \text{ and } T_j \text{ occur})}{P(S_i \text{ occurs})}$$

$$=\frac{P(S_i\cap T_j)}{P(S_i)},$$

which, when rearranged, yielded the following equation for the joint probabilities:

$$P(S_i\cap T_j)=P(T_j|S_i)P(S_i).$$

Thus, the joint probabilities were found by taking the matrix product of Tables 2 and 5. A complete summary of these joint probabilities is given in Table 6.

Table 6 also reflects, along the bottom row, the marginal probability associated with each test outcome. These probabilities were computed using the equation

$$P(T_j)=P(T_j|S_1)P(S_1)+P(T_j|S_2)P(S_2)+P(T_j|S_3)P(S_3)+P(T_j|S_4)P(S_4),$$

which, in effect, is a summation of the column entries in Table 6.

TABLE 6. *Joint Probabilities That Both the Test Outcomes and the True States of Modification Improvement Occur* $P(S_i \cap T_j)$

STATE	TEST OUTCOMES				$P(S_i)$
	T_1	T_2	T_3	T_4	
S_1	0.2250	0.0125	0.0063	0.0062	0.2500
S_2	0.0250	0.4375	0.0250	0.0125	0.5000
S_3	0.0050	0.0100	0.1750	0.0100	0.2000
S_4	0.0013	0.0012	0.0025	0.0450	0.0500
$P(T_j)$	0.2563	0.4612	0.2088	0.0737	1.000

The final step in computing the relevant probabilities was to find the posterior probability associated with each level of improvement, conditioned on the test outcomes. The conditional relationship used to find these posterior probabilities was

$$P(S_i|T_j)=\frac{P(S_i \cap T_j)}{P(T_j)}.$$

A complete summary of these posterior probabilities is given in Table 7.

TABLE 7. *Posterior Probabilities of Achieving Stated Levels of Modification Improvement Conditioned on Test Outcomes*

$P(S_i|T_j)$

STATE	TEST OUTCOMES			
	T_1	T_2	T_3	T_4
S_1	0.8779	0.0271	0.0302	0.0841
S_2	0.0975	0.9486	0.1197	0.1696
S_3	0.0195	0.0217	0.8381	0.1357
S_4	0.0051	0.0026	0.0120	0.6106

EXPECTED SYSTEM COSTS CONDITIONED ON TEST RESULTS.

The next step in the decision procedure was to use the probabilities in Table 7 to compute the expected 3½-year system cost associated with each test outcome. These costs were computed using the equation

$$E(C^{**}|T_j)=P(S_1|T_j)PV_1+P(S_2|T_j)PV_2+$$
$$P(S_3|T_j)PV_3+P(S_4|T_j)PV_4.$$

Summing these expected costs for all four levels of improvement, then, gave the expected cost associated with each test outcome. That is,

$$\begin{aligned} E(C^{**}|T_1) &= (0.8779)(38.04)+(0.0975)(47.52)+\\ &\quad (0.0195)(57.01)+(0.0051)(66.49)\\ &= 33.40+4.63+1.11+0.34\\ &= \$39.48 \text{ million},\end{aligned}$$

$$\begin{aligned} E(C^{**}|T_2) &= (0.0271)(38.04)+(0.9486)(47.52)+\\ &\quad (0.0217)(57.01)+(0.0026)(66.49)\\ &= 1.03+45.08+1.24+0.17\\ &= \$47.52 \text{ million},\end{aligned}$$

$$\begin{aligned} E(C^{**}|T_3) &= (0.0302)(38.04)+(0.1197)(47.52)+\\ &\quad (0.8381)(57.01+(0.0120)(66.49)\\ &= 1.15+5.69+47.78+0.80\\ &= \$55.42 \text{ million},\end{aligned}$$

and

$$\begin{aligned} E(C^{**}|T_4) &= (0.0841)(38.04)+(0.1696)(47.52)+\\ &\quad (0.1357)(57.01)+(0.6106)(66.49)\\ &= 3.20+8.06+7.74+40.60\\ &= \$59.60 \text{ million}.\end{aligned}$$

The expected system cost associated with each test outcome was then used, in conjunction with the marginal probability of each test outcome, to compute the expected system cost, given only that testing was conducted. This cost was computed using the equation

$$\begin{aligned} E(C^{**}|\text{TEST}) &= P(T_1)\,E(C^{**}|T_1)+P(T_2)\,E(C^{**}|T_2)+\\ &\quad P(T_3)\,E(C^{**}|T_3)+P(T_4)\,E(C^{**}|T_4).\end{aligned}$$

Summing these expected costs for all four test outcomes, then, gave the expected system cost conditioned only on the test being conducted. That is,

$$\begin{aligned} E(C^{**}|\text{TEST}) &= (0.2563)(39.48)+(0.4612)(47.52)+\\ &\quad (0.2088)(55.42)+(0.0737)(56.96)\\ &= 10.12+21.92+11.57+4.20\\ &= \$47.81 \text{ million}.\end{aligned}$$

Note that if the test result turned out to be T_4, the decisionmaker would choose not to apply the modification, since the system cost with modification (\$59. 60 mil) would exceed the system cost without modification (\$56. 96 mil). This, in essence, constituted the return from the testing effort.

PERMISSIBLE EXPENDITURES ON FURTHER DEVELOPMENT AND TESTING.

As indicated earlier in the paper, the major question faced by the decisionmaker was: How much was it economically permissible to spend on further development and testing? The answer to this question was computed in two parts.

First, the maximum level that could be spent on further development only (without testing), denoted A_D, was computed as the difference between the expected system cost if the modification was applied without improvement, $E(C^*)$, and the expected system cost if the modification was improved but not tested before application, $E(C^{**})$. That is,

$$\begin{aligned} A_D &= E(C^*) - E(C^{**}) \\ &= 53.29 - 48.00 \\ &= \$5.29 \text{ million.} \end{aligned}$$

Next, the maximum level that could be spent on testing undertaken in conjuction with further development, denoted A_T, was computed as the difference between expected system cost with improvement but without testing and expected system cost with both improvement and testing. That is,

$$\begin{aligned} A_T &= E(C^{**}) - E(C^{**}|TEST) \\ &= 48.00 - 47.81 \\ &= \$0.19 \text{ million.} \end{aligned}$$

Therefore, if testing was undertaken in conjunction with an improvement program, it was economical to spend an amount on testing not to exceed $190,000.

TESTING WITHOUT FURTHER DEVELOPMENT.

It was also desired to evaluate the alternative of testing without further development. For this alternative it was assumed that two months would be needed for the test to be completed and the modification procedure applied. Therefore, the system costs were estimated to be distributed over the 3½-year period as indicated in Table 8.

TABLE 8. *System Costs for Testing Without Further Development* ($ MILLION)

STATE (S_i)	INITIAL COST	PERIODIC REPAIR COSTS ‡					PRESENT VALUE (PV_i)
		1	2	3	4	5	
S_1	0 *	13.33 †	2.67	8	8	8	34.63
S_2	0 *	13.33 †	4.00	12	12	12	45.39
S_3	0 *	13.33 †	5.33	16	16	16	56.14
S_4	0 *	13.33 †	6.67	20	20	20	66.90

*Cost of testing not included. The permissible expenditure for this effort is a model output.

†Includes two-month repair cost with modification plus $10 million for cost of modification application.

‡First period is two months; second period is four months; others are 12 months.

The expected system cost for this alternative for each test outcome was given by the equation

$$\begin{aligned} E(C^*|T_j) = {} & P(S_1|T_j)PV_1 + P(S_2|T_j)PV_2 + \\ & P(S_3|T_j)PV_3 + P(S_4|T_j)PV_4. \end{aligned}$$

The actual computations of the posterior probabilities, $P(S_i|T_j)$, for this alternative are not shown; however, they were based on the prior probabilities listed in Table 1 rather than those listed in Table 2. Furthermore, the discounted costs, PV_i, used were those listed in Table 8. Thus, as before,

$$\begin{aligned} E(C^*|T_1) &= (0.7660)(34.63)+(0.0851)(45.39)+ \\ &\quad (0.1064)(56.14)+(0.0426)(66.90) \\ &= 26.53+3.86+5.97+2.85 \\ &= \$39.21 \text{ million}, \end{aligned}$$

$$\begin{aligned} E(C^*|T_2) &= (0.0238)(34.63)+(0.8333)(45.39)+ \\ &\quad (0.1190)(56.14)+(0.0238)(66.90) \\ &= 0.83+37.82+6.68+1.\ 59 \\ &= \$46.92 \text{ million}, \end{aligned}$$

$$\begin{aligned} E(C^*|T_3) &= (0.0054)(34.63)+(0.0217)(45.39)+ \\ &\quad (0.9511)(56.14)+(0.0217)(66.90) \\ &= 0.19+0.99+53.39+1.45 \\ &= \$56.02 \text{ million}, \end{aligned}$$

and

$$\begin{aligned} E(C^*|T_4) &= (0.0118)(34.63)+(0.0235)(45.39)+ \\ &\quad (0.1176)(56.14)+(0.8471)(66.90) \\ &= 0.41+1.07+6.60+56.67 \\ &= \$64.75 \text{ million}. \end{aligned}$$

The expected system cost associated with each test outcome was then used to compute the expected system cost given only that testing was conducted. That is,

$$\begin{aligned} E(C^*|\text{TEST}) = {} & P(T_1)\,E(C^*|T_1)+P(T_2)\,E(C^*|T_2)+ \\ & P(T_3)\,E(C^*|T_3)+P(T_4)\,E(C^*|T_4), \end{aligned}$$

where, again, the marginal probabilities associated with the test outcomes were computed based on the prior probabilities listed in Table 1. Thus,

$$\begin{aligned} E(C^*|\text{TEST}) &= (0.1175)(39.21)+(0.2100)(46.92)+ \\ &\quad (0.4600)(56.02)+(0.2125)(56.96) \\ &= 4.61+9.85+25.77+12.11 \\ &= \$52.34 \text{ million}. \end{aligned}$$

Finally, the maximum level that could be invested in testing without further development, $\hat{A}_T$, was computed as follows:

$$\begin{aligned} \hat{A}_T &= E(C^*)-E(C^*|\text{TEST}) \\ &= 53.29-52.34 \\ &= \$0.95 \text{ million}. \end{aligned}$$

Therefore, even without further development, it was economical to spend an amount on testing not to exceed $950,000 merely to reduce the uncertainty surrounding the mean prior estimate of failure rate reduction.

DECISION TREE

The foregoing procedure may also be portrayed in decision tree format. The decision tree, complete with expected costs and the probabilities associated with those costs, is presented in Figure 1.

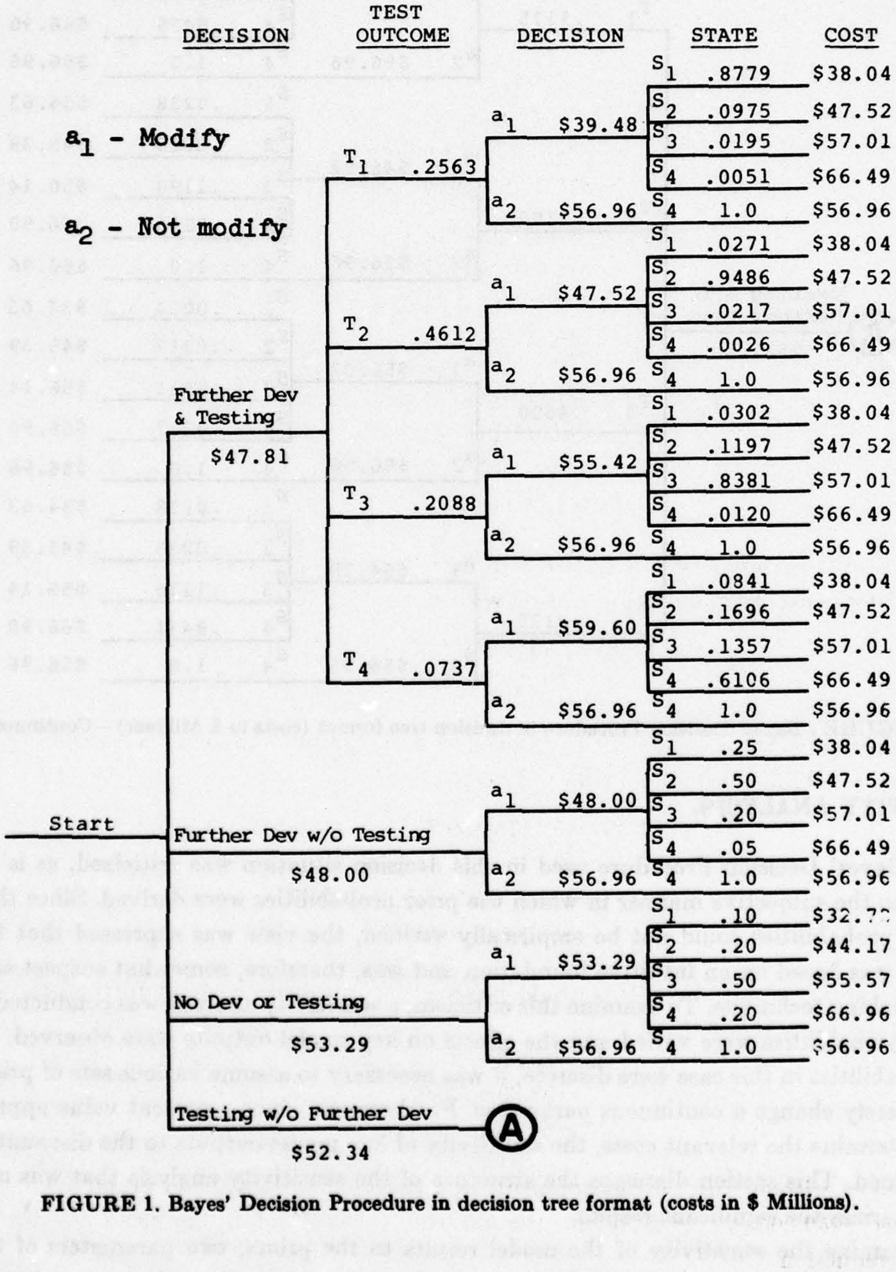


FIGURE 1. Bayes' Decision Procedure in decision tree format (costs in $ Millions).

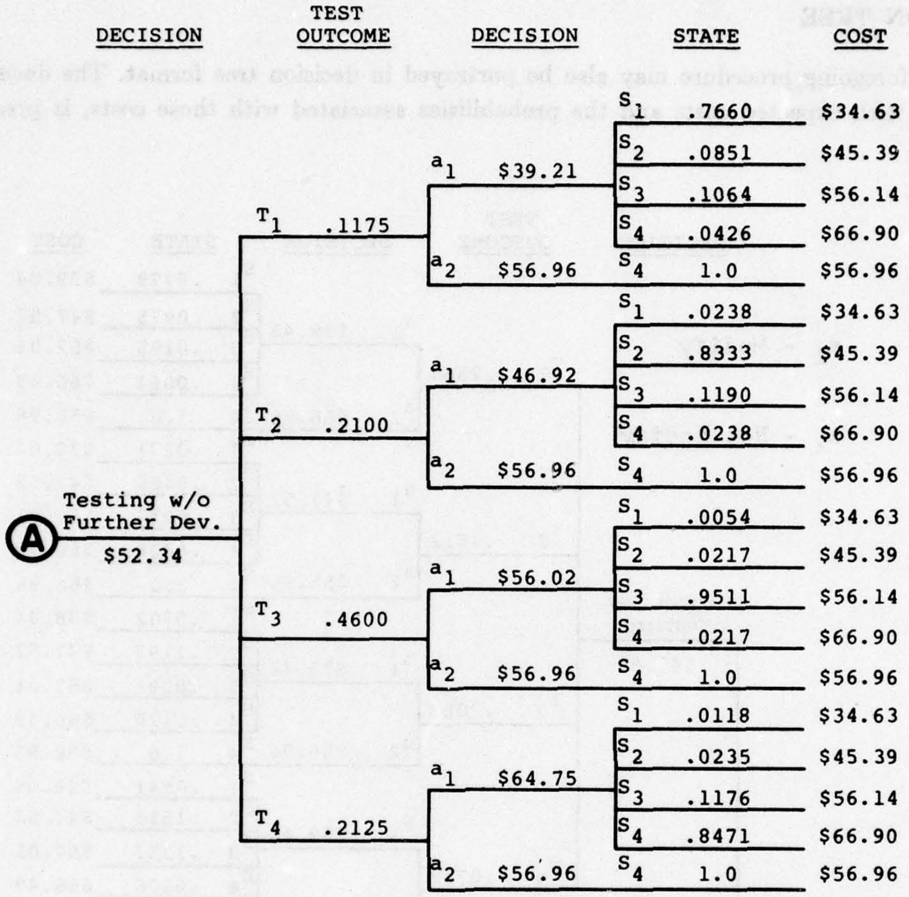


FIGURE . Bayes' Decision Procedure in decision tree format (costs in $ Millions) — Continued.

SENSITIVITY ANALYSIS.

The Bayes' Decision Procedure used in this decision situation was criticized, as is often the case, due to the subjective manner in which the prior probabilities were derived. Since the subjective prior probabilities could not be empirically verified, the view was expressed that the entire procedure was based on an intuitive foundation and was, therefore, somewhat suspect as a useful decision-making technique. To examine this criticism, a sensitivity analysis was conducted in which the prior probabilities were varied and the effects on key model outputs were observed. Since the prior probabilities in this case were discrete, it was necessary to assume various sets of priors rather than to merely change a continuous parameter. Furthermore, since a present value approach was used to determine the relevant costs, the sensitivity of key model outputs to the discount rate was also examined. This section discusses the structure of the sensitivity analysis that was conducted and summarizes the significant results.

To examine the sensitivity of the model results to the priors, two parameters of the priors were used—the arithmetic mean and the variance about the mean. The prior mean was a measure of the tendency of the group of estimators toward a central value and, in a sense, represented the collective opinion of the group as to the true value of the parameter. Similarly, the variance indicated the spread of the priors about the mean and was a measure of the uncertainty in the group concerning the true value of the parameter that the group members were trying to estimate. More specifically, the mean prior estimate could logically be said to reflect the collective knowledge of the members of the group concerning the physical system under consideration; the variance could be said to reflect the collective strength of their convictions about the true parameter value.

TABLE 9. *Data Sets for Sensitivity Analysis*

STATE (S_i)	SET #1	SET #2	SET #3	SET #4	SET #5
CONSTANT VARIANCE (260) (%)^2					
S_1	0.10	0.10	0.25	0.40	0.55
S_2	0.15	0.45	0.50	0.45	0.35
S_3	0.60	0.35	0.20	0.10	0.05
S_4	0.15	0.10	0.05	0.05	0.05
MEAN (%)	24	31	39	44	48
CONSTANT MEAN (39) (%)					
S_1	0.15	0.20	0.25	0.25	0.30
S_2	0.70	0.60	0.50	0.55	0.45
S_3	0.10	0.15	0.20	0.10	0.15
S_4	0.05	0.05	0.05	0.10	0.10
VAR (%)^2	179	219	260	299	339

In choosing the sets of prior probabilities to examine, we selected five sets with a constant mean and five sets with constant variance. The sets of prior probabilities used are set forth in Table 9. The two groups of prior probability sets had one set, the base estimate, in common. Thus, by solving the model for each of the prior probability groups, it was possible to separately examine the sensitivities of model outputs to the prior mean, holding variance constant, and to prior variance holding the mean value constant.

The first outputs examined for sensitivity were the principal model outputs—permissible expenditure level on modification development, A_D, and permissible expenditure level on testing undertaken in conjunction with modification development, A_T. These two outputs were examined for sensitivity to both prior mean and prior variance and separately to the discount rate. In addition, the expected system costs for the five alternatives considered were examined for sensitivity to the discount rate.

The results relating A_D to prior mean and variance are illustrated in Figure 2. Two observations based on these results are particularly noteworthy.

First, A_D was found to be an increasing linear function of the prior mean. In fact, since the prior mean could be viewed as representing the reliability of the modification procedure, the implication was that an increasing linear relationship existed between reliability and the level of expenditure permissible for modification improvement. Moreover, an examination of the slope of the linear relationship revealed that a change in reliability of one percent would cause an average change of $474,200 in A_D. Thus, the relationship was not only linear but quite sensitive, indicating that further efforts to improve the reliability of the modification procedure might well pay significant dividends. These results are analogous to those reported by Brown and Perlman [1] in their study describing a repair parts inventory model for the F-14 aircraft.

Second, A_D was found to be totally unrelated to the prior variance. That is, the level of uncertainty exhibited in the prior probabilities, by way of variance about the mean, had nothing at all to do with determining the level of permissible expenditures on modification development—which

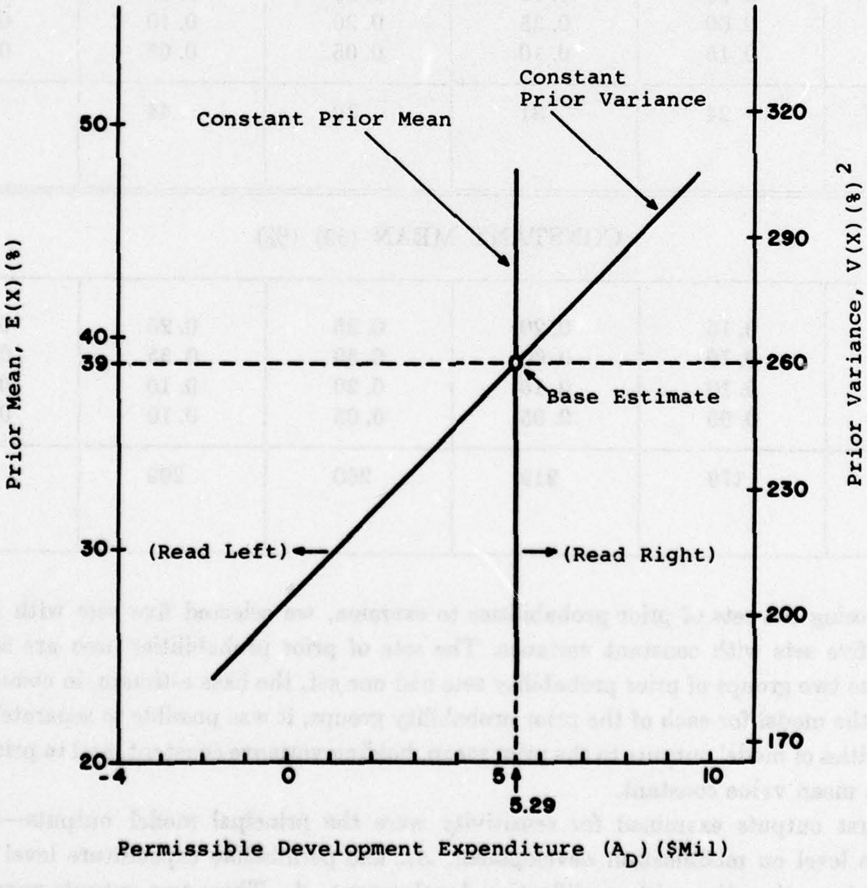


FIGURE 2. Influence of prior mean and variance on permissible expenditure level for modification development.

for a prior mean of 39 percent, was a constant $5.29 million. That is to say, with respect to A_D, it was the collective knowledge of the group of engineers estimating the priors (exhibited in their tendency toward a central value) that was important, rather than the strength of the group's conviction about the mean value. Although it is difficult to completely divorce a discussion of means from a discussion of variances, the relationships exhibited by Fig. 2 would seem to support a general hypothesis that the value of priors in decision-making is positively correlated to the professional competence of the group estimating the priors, and that the uncertainty in the group is not as important as bias or the lack of objectivity on the part of the group members.

The results relating A_T to prior mean and variance are illustrated in Figure 3. The relationships exhibited by Fig. 3 were perhaps the most germane of the sensitivity analysis since they dealt more directly with the principal issue of the analysis—that of acquiring additional information through testing. Two major observations were important in this respect.

First, unlike permissible development expenditures, permissible testing expenditures varied with both prior mean and prior variance. As prior variance increased, holding the mean constant, the permissible expenditures on testing increased at an increasing rate. This implied that the higher the level of uncertainty in the prior probabilities, the greater the payoff from testing was

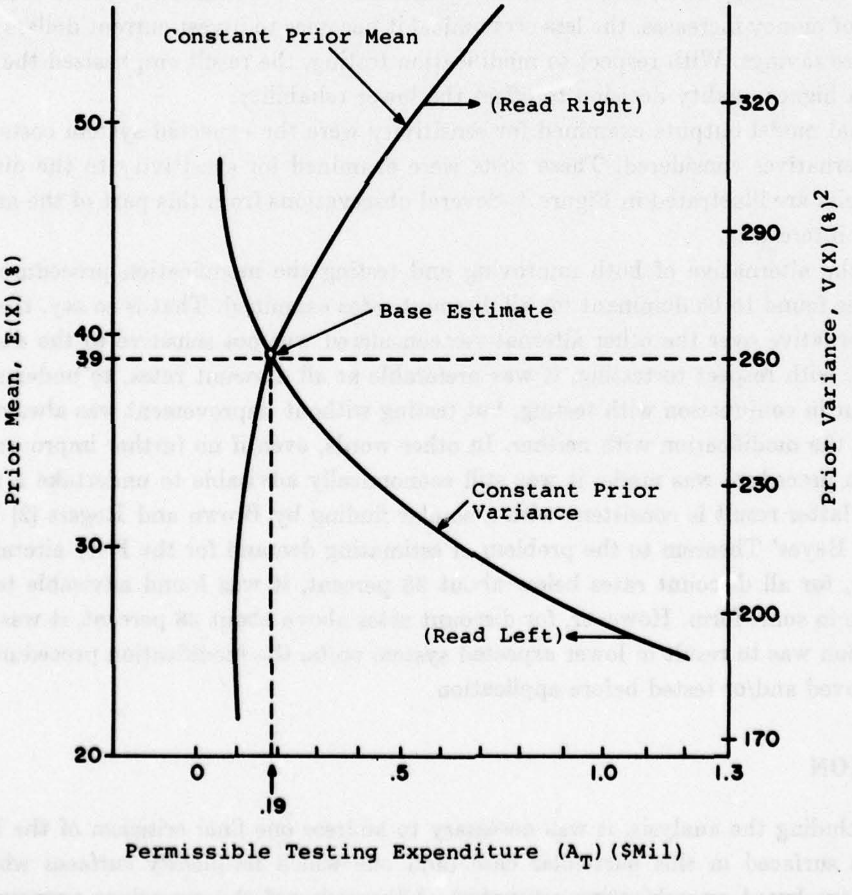


FIGURE 3. Influence of prior mean and variance on permissible expenditure level for modification testing.

likely to be. On the other hand, as the prior mean increased, holding the variance constant, the permissible expenditures on testing decreased at a decreasing rate. This implied that the more optimistic the prior estimate was concerning the level of reduction in failure rate expected to be brought about by modification improvement, the smaller the expected payoff from testing was likely to be.

Second, a major implication of Fig. 3 was that prior estimates should conform as closely as possible to the true results being estimated. Stated in another way, it was important that the prior estimates represent the most informed judgment available. Despite this obvious finding, however, it was instructive to observe that regardless of the accuracy and uncertainty characterizing the prior estimates, the Bayes' Decision Procedure tended to produce a higher quality decision. This latter observation was deduced from the result that, for all values of the prior mean and variance examined, it was advisable to invest in testing to secure additional information before making a modification decision.

The model outputs, A_D and A_T, were also examined for sensitivity to the discount rate. However, the only significant finding in this respect was that as the discount rate increased, there was a tendency to spend less on product improvement and more on product testing. With respect to modification development, this result merely illustrated the well-known fact that as the opportunity cost of money increases, the less economical it becomes to invest current dollars in anticipation of future savings. With respect to modification testing, the result emphasized the importance of making a higher quality decision to offset the lower reliability.

The final model outputs examined for sensitivity were the expected system costs for each of the five alternatives considered. These costs were examined for sensitivity to the discount rate, and the results are illustrated in Figure 4. Several observations from this part of the analysis were particularly interesting.

First, the alternative of both improving and testing the modification procedure before application was found to be dominant for all discount rates examined. That is to say, the preference for this alternative over the other alternatives considered was not sensitive to the discount rate.

Second, with respect to testing, it was preferable at all discount rates, to undertake product improvement in conjunction with testing, but testing without improvement was always preferable to applying the modification with neither. In other words, even if no further improvement in the modification procedure was made, it was still economically advisable to undertake a testing program. This latter result is consistent with a similar finding by Brown and Rogers [2] in their application of Bayes' Theorem to the problem of estimating demand for the F–14 aircraft.

Finally, for all discount rates below about 28 percent, it was found advisable to apply the modification in some form. However, for discount rates above about 28 percent, it was found that if modification was to result in lower expected system costs, the modification procedure had to be either improved and/or tested before application.

CONCLUSION

In concluding the analysis, it was necessary to address one final criticism of the Bayes' Procedure that surfaced in this particular case (and one which frequently surfaces when decision procedures are based on subjective estimates). Adversaries of the procedure suggested, as they often do, that if the prior estimates were overly optimistic, the result would be higher—not lower—

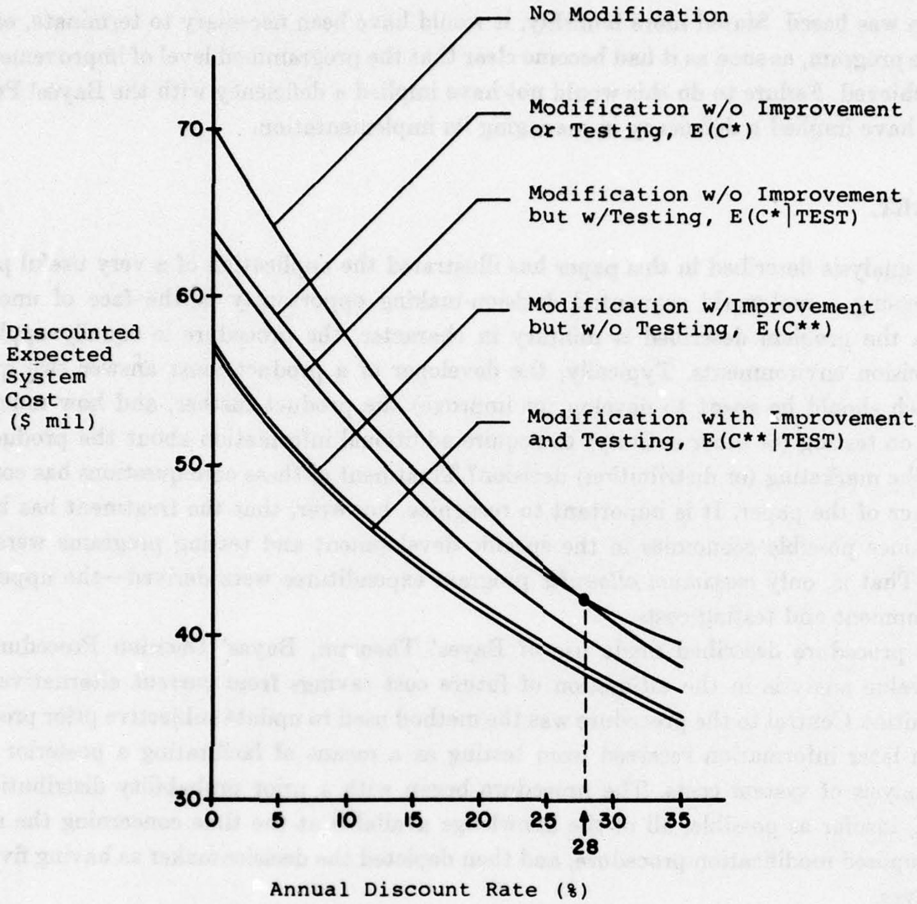


FIGURE 4. Influence of annual discount rate on discounted expected system cost.

net system cost. The rationale for this contention may be seen in the following discussion related to the problem addressed in this paper:

> **Recall that an expected failure rate reduction level of 24 percent could be achieved by applying the modification without further development, and that further development was expected to improve this reduction level to 39 percent. It follows, then, that the permissible investment level in further development (A_D) of $5.29 million was based on achieving an increase in the mean failure rate reduction of not less than 15 percent. If, after spending the $5.29 million, it had been found that a failure rate reduction of less than 15 percent had been achieved, the conclusion could have been justifiably drawn that it would have been cheaper to have applied the modification without further development. This was the danger, the critics claimed, in basing a technique on subjectively determined prior probabilities that reflected the human characteristic of being overly optimistic.**

How does one handle this apparent anomaly? In a general sense, the answer to this question constitutes one of the key requirements in making the Bayes' Procedure work. In this particular case, the answer was that the amount spent on development *could not* be permitted to exceed the programmed amount. If it had, the critics would have been correct—the procedure would have been more expensive. Therefore, it was necessary that the funds allocated to product improvement be spent in a manner that achieved the *programmed* level of improvement upon which the fund

allocation was based. Stated more bluntly, it would have been necessary to terminate, or at least revise the program, as soon as it had become clear that the programmed level of improvement would not be achieved. Failure to do this would not have implied a deficiency with the Bayes' Procedure; it would have implied a deficiency in managing its implementation.

SUMMARY.

The analysis described in this paper has illustrated the application of a very useful procedure for addressing a real-world sequential decision-making opportunity in the face of uncertainty. Although the problem described is military in character, the procedure is equally applicable in other decision environments. Typically, the developer of a product must answer two questions: How much should be spent to develop (or improve) the product further, and how much should be spent on testing (or other activity) to acquire additional information about the product before making the marketing (or distribution) decision? Treatment of these core questions has constituted the essence of the paper. It is important to recognize, however, that the treatment has been only partial, since possible economies in the specific development and testing programs were not addressed. That is, only *maximum allowable* program expenditures were derived—the upper bounds on development and testing costs.

The procedure described made use of Bayes' Theorem, Bayes' Decision Procedure, and a present-value analysis in the estimation of future cost savings from current alternative decision opportunities. Central to the procedure was the method used to update subjective prior probabilities based on later information received from testing as a means of facilitating a posterior expected value analysis of system costs. The procedure began with a prior probability distribution which captured, insofar as possible, all of the knowledge available at the time concerning the reliability of the proposed modification procedure, and then depicted the decisionmaker as having five distinct alternatives.

First, he could choose not to apply the modification and to continue to sustain the present high system costs. It was shown that this alternative was not economically advisable, since application of the modification even without further improvement would result in a lower expected system cost.

Second, he could choose to apply the modification without further improvement or testing. However, based on the prior probabilities given, it was shown that this alternative could also be improved on by developing the modification procedure further or by testing it without further development before application.

Third, he could choose to improve the modification and then apply it without first subjecting it to a testing procedure. However, it was shown that testing would give information to the decisionmaker that he could use to improve the quality of his decision and further reduce the expected system cost.

Fourth, he could choose to test the modification procedure without improvement before application. Although this alternative was preferable to the alternatives of either not applying the modification or applying it without testing, it was shown that this alternative could also be improved on by developing the modification procedure further before application, and that this relationship held for all annual discount rates up to and including 35 percent.

Finally, the fifth alternative—that of improvement and testing before application—was shown to be the most economical alternative based on expected system cost over a 3½-year period.

A comprehensive sensitivity analysis was then conducted to examine the sensitivity of the key model outputs to both the mean and variance of the prior probabilities and separately to the discount rate. Several of the results of the sensitivity analysis were significant. They showed, for example, that

1. The permissible amount to spend on further development of the modification procedure was a sensitive increasing linear function of the prior mean but was totally unrelated to the prior variance. That is to say, with respect to this particular output, it was the collective knowledge of the group of engineers estimating the priors (exhibited in their tendency toward a central value) that was important, rather than the strength of the group's conviction (prior variance) about the mean value.

2. The permissible amount to spend on testing was sensitive to both the prior mean and prior variance. The implication of this was that the prior estimates should represent the most informed judgment available. However, since testing expenditures were found to be permissible over a wide range of prior probabilities, it was also clear that the Bayes' Decision Procedure could result in a higher quality decision even when the prior estimates represented a low order of prior knowledge.

3. The permissible amount to spend on further development varied inversely with the discount rate, and the permissible amount to spend on testing varied directly with the discount rate. That is to say, as the discount rate increased, the tendency was to spend less on product improvement and more on product testing.

4. The alternative of both improving and testing the modification procedure before application dominated the other four alternatives for all discount rates examined (0% to 35%). In addition, the alternative of testing the modification procedure without further improvement before application dominated the alternative of neither improving nor testing the procedure. This later finding was particularly interesting, since it indicated that expected system cost could be lowered by investing only in efforts to reduce uncertainty. That is, even if no further improvement in the modification procedure was made, it was still economically advisable to undertake a testing program.

The results of the analysis indicated that both improvement and testing should be undertaken and that the maximum amounts that should be spent on the programs were $5.29 million and $0.19 million, respectively. As indicated previously, the actual amount to spend on each program was not addressed since that would have involved an examination of the economies exhibited by the specific development and testing programs used, which was not done until later. However, the analysis showed that to the extent it was possible to conduct the development and testing programs for amounts less than the maximums permissible, substantial savings could be realized by postponing application of the procedure to permit it to be improved and tested.

The results of the analysis stimulated management in at least two ways: first, to secure the most reputable engineers available to estimate the priors, and then to undertake both development and testing as described in alternative five. The development and testing programs were actually conducted for amounts less than the maximums permissible and, as a result, sizeable savings were realized.

REFERENCES

[1] Brown, George F. and Bernard L. Perlman, "Optimal Inventory Management for Naval Aviation Support," Center for Naval Analysis, Research Contribution 186 (1971).

[2] Brown, George F. and Warren F. Rogers, "A Bayesian Approach to Demand Estimation and Inventory Provisioning," Naval Research Logistics Quarterly, (December 1973).

[3] Raiffa, Howard, *Decision Analysis*, (Addison-Wesley, Reading, Mass., 1970).

[4] Circular A-94, "Discount Rates to be used in Evaluating Time-Distributed Costs and Benefits," Office of Management and Budget, Washington, D.C., (27 March 1972).

THE CYCLIC SEPARATION SCHEDULING PROBLEM*

Ronald D. Armstrong

The University of Texas at Austin
Austin, Texas

Prabhakant Sinha
University of Massachusetts
Amherst, Massachusetts

ABSTRACT

This paper deals with the problem of scheduling items (tasks, employees, equipment, etc.) over a finite time horizon so as to minimize total cost expenditures while maintaining a predefined separation between certain items. The problem is cyclic, because the same schedule will be repeated over several consecutive time periods of equal length. Thus, requirements are present to maintain the separation of items not only within the individual time periods considered, but also between items in adjoining periods. A special purpose branch-and-bound algorithm is developed to solve this scheduling problem by taking advantage of its cyclic nature. Computational results are given.

1. INTRODUCTION

Through the years, the classical assignment problem has had numerous offspring. These include the multidimensional assignment problem [10], the quadratic assignment problem [6, 7, 9], the traveling salesman problem [8], and the generalized assignment problem [11]. This paper adds the cyclic separation scheduling problem to the list of progeny.

The Separation Scheduling Problem (SSP) consists of assigning items (tasks, employees, equipment, etc.) in such a way that a predefined distance is maintained between certain items. We define S to be the separation matrix, where an element $s_{jq}=s_{qj}$ specifies that item j must be separated from item q by s_{jq} positions. For the purpose of establishing terminology and to lead to the application where our work is currently being directed, we assume that n items are available for scheduling on m days. One and only one item may be assigned on any day, and the number of times an item may be scheduled during the m days is limited. When the objective is to minimize the total cost of assigning the items, the separation scheduling problem can be formulated as the following integer programming problem:

*This study was supported, in part, by the Office of Naval Research.

$$\text{Minimize } \sum_{i=1}^{m} \sum_{j=1}^{n} c_{ij} x_{ij}, \tag{1}$$

subject to

$$\sum_{j=1}^{n} x_{ij} = 1, \qquad i = 1, 2, \ldots, m; \tag{2}$$

$$\sum_{i=1}^{m} x_{ij} \leq u_j, \qquad j = 1, 2, \ldots, n; \tag{3}$$

$$x_{ij} + x_{rq} \leq 1 \quad \text{if } s_{jq} > d_{ir}, \quad \begin{array}{l} r = 1, 2, \ldots, m-1; \\ i = r+1, r+2, \ldots, m; \\ j = 1, 2, \ldots, n; \\ q = 1, 2, \ldots, n; \end{array} \tag{4}$$

$$x_{ij} = 0 \text{ or } 1 \qquad \begin{array}{l} i = 1, 2, \ldots, m; \\ j = 1, 2, \ldots, n, \end{array} \tag{5}$$

where

$$x_{ij} = \begin{cases} 1, & \text{if item } j \text{ is scheduled on day } i \\ 0, & \text{otherwise,} \end{cases}$$

$d_{ir} = i - r$, c_{ij} is the cost of assigning item j on day i, and u_j is the maximum number of times item j may appear during m days.

This problem was formulated and solved by Balintfy [2] by using an algorithm similar to that of Little, et al. [8] to solve the traveling salesman problem. The problem occurs in the context of scheduling menu items over the days of the planning horizon, where the purpose of separation constraints is to ensure variety in the schedule. It is conceivable, however, that separation scheduling has the opposite goal—to ensure proximity of certain items within specified levels. The modifications to our algorithm to account for the latter aspect will be discussed in the conclusions.

In menu plans [1], if a schedule covering m days is thought of as being repeated every m days, the item allocated on day 1 will be the same as the item on day $m+1$, the item on day 2 will be the same as that on day $m+2$, and so on. This cyclic property implies that, if item A is constrained to be separated from item B by at least two days, and item A appears on day 1 of the schedule, then item B should not appear, not only on days 2 and 3, but also on days m and $m-1$. Such a schedule can be conceptualized as the allocation of items to equispaced points on the circumference of a circle, and the separation between two points is then the minimum of the distances between them in the clockwise and counterclockwise directions. More formally, d_{ir}, the distance between discrete points i and r, is defined by

$$d_{ir} = \min\{(i-r), m-(i-r)\}. \tag{6}$$

An additional feature of the problem is that the cost of an item depends only on the item, and not on where it is scheduled. Thus, if a cost matrix C is defined such that c_{ij} is the cost of assigning item j to day i, the rows of the cost matrix are identical; i.e., $c_{ij} = c_{rj}$ for all i, r, j. It is this property, along with the cyclic nature of the separation constraints, which is used to defined the Cyclic Separation Scheduling problem (CSSP):

$$\text{(7)} \qquad \text{Minimize } \sum_{j=1}^{n} \sum_{i=1}^{m} c_j x_{ij} \text{ subject to (2)—(5),}$$

where the distance function is now defined by (6), and the separation matrix S is defined as before.

The algorithm of Balintfy used to solve the SSP applies directly to the CSSP, as the CSSP is only a special case of the SSP. However, no advantage is taken of the cyclic property of the separation constraints and the special structure of the cost matrix. The objective addressed here is to exploit the aforementioned properties in developing a more efficient solution technique.

2. SEPARATION SCHEDULING

The fundamental features of the parent branch-and-bound algorithm developed and implemented by Balintfy are recapitulated in this section.

A branch-and-bound algorithm involves breaking up the set of all solutions into smaller and smaller subsets and finding lower bounds on the costs of the subsets until a single solution is found with a cost less than or equal to the lower bounds on the costs of all other subsets of solutions. Consequently, in any branch-and-bound algorithm, a set of rules is required for (a) branching, i.e., partitioning solution sets, (b) determining lower bounds on sets of solutions, (c) selecting the set of solutions to partition next, and (d) recognizing infeasible, suboptimal, or optimal solutions.

In the case of the separation scheduling problem, a feasible solution is represented by a set of ordered pairs: $\{(i_1, j_1), (i_2, j_2), \ldots, (i_m, j_m)\}$. The inclusion of pair (i, j) in this set means that item j is scheduled to appear in time period i. A set of solutions will be represented by $W(S)$, where S is a set of ordered pairs characterizing the solution set, and consists of (a) a subset of ordered pairs $\underline{S}$ representing *allocations*, and (b) a subset of ordered pairs $\bar{S}$ representing *prohibitions*. If $\underline{(i, j)}$ is in $\underline{S}$, this means that every solution in $W(S)$ must have item j scheduled on day i, and if $\overline{(i, j)}$ is in $\overline{S}$, item j must be barred from day i in every solution in $W(S)$. The bars are placed under or over the ordered pair to indicate which part of S the element belongs, to. When no bar appears, the ordered pairs referred to may be either from $\underline{S}$ or $\overline{S}$.

In the separation scheduling algorithm, to partition solution sets, an ordered pair (i, j) is selected as the *partitioning element* such that for the solution set $W(S)$ being partitioned, (i, j) is not in S. Then $W(S_1)$ and $W(S_2)$ constitute the partition of $W(S)$, where $S_1 = SU(\overline{i, j})$ and $S_2 = SU(\overline{i, j})$. Thus, $W(S_1)$ represents the subset of solutions from $W(S)$ which have item j barred from being on the ith day, and $W(S_2)$ represents the subset of solutions of $W(S)$ which have item j scheduled on the ith day. This method of partitioning is similar to the algorithm of Little, et al. to solve the traveling-salesman problem.

The process of branching is conveniently represented by a tree, with the nodes representing sets of solutions. A branch connects node N and N_1 of the tree when the solution set $W(S)$ represented by N contains the solution set $W(S_1)$ represented by N_1. $L(N)$ will then denote a lower bound on the value of the objective function for all the solutions represented by node N. No further partitioning is required at node N when $L(N)$ is greater than the objective function value associated with any known feasible schedule. In such a case, the node is said to be fathomed. Fathoming also takes place when either a feasible schedule is achieved or it is ascertained that no feasible solution can be found in $W(S)$.

The features of the SSP which make it a restricted assignment problem are the separation constraints. These constraints are utilized in the algorithm to maintain a separation between certain items. No allocations are permitted which would violate the constraints of the problem.

A statement of the algorithm that contains a detailed description of the rules used in our implementation is given in section 4.

3. CYCLIC SEPARATION SCHEDULING PROBLEM

In branch-and-bound algorithms, the terminal nodes of the solution tree represent the set of all solutions. In the CSSP, very early in the branching process it is possible to eliminate whole subsets of solutions from consideration. With the cyclic property of the problem in mind, the following definitions are given.

DEFINITION: A schedule X' is a *translation* of a schedule X^* if $x'_{i\oplus k,j}=x^*_{ij}$, $i=1, 2, \ldots, m$; $j=1, 2, \ldots, n$, for some constant integer k, $0\leq k\leq m$, where

$$i\oplus k=\begin{cases} i+k, & i+k\leq m \\ i+k-m, & i+k>m. \end{cases} \tag{8}$$

DEFINITION: A schedule X' is a *mirror image* of a schedule X^* about position k if $x^*_{i\oplus k,j}=x'_{k\theta i,j}$, $i=0, 1, 2, \ldots, m-1$; $j=1, 2, \ldots, n$; where $i \oplus k$ is defined by (8), and

$$k\theta i=\begin{cases} k-i, & k-i>0 \\ m+(k-i), & k-i\leq 0. \end{cases}$$

Consider, for example, the following three solutions to a seven-day CSSP:

Day	Solution 1	Solution 2	Solution 3
1	A	D	B
2	B	A	A
3	A	B	D
4	C	A	C
5	D	B	A
6	A	A	B
7	B	C	A

Notice that Solution 2 can be obtained from Solution 1 by shifting all the allocations three days ahead. Thus, Solution 2 is a translation of Solution 1. Also, Solution 3 can be obtained from Solution 1 by taking a mirror image of the allocations about day 4.

DEFINITION: Two schedules X' and X^* are *equivalent* if either X^* is a translation of X', or a third schedule exists that is a translation of X' and a mirror image of X^*.

Thus, Solutions 1, 2, and 3 listed above are all equivalent. When showing the equivalence between Solution 1 and Solution 2, the third solution mentioned in the definition is given by $k=0$. In other words, the trivial translation gives back Solution 1.

This definition of equivalence among solution sets preserves (a) the frequency of each item on the schedule, and (b) the separation in days between item pairs. Therefore, infeasibiliy, suboptimality, or optimality of one implies infeasibility, suboptimality, or optimality of the other, respectively.

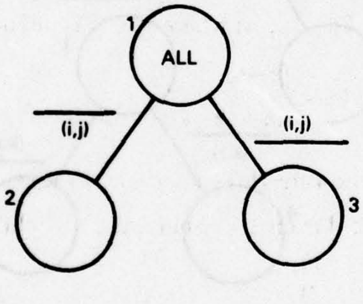
DEFINITION: A *dormant* node is a node corresponding to a solution set characterized by prohibitions only.

In terms of the solution tree, a dormant node is one which is connected to the start of the tree through branches corresponding to prohibitions only. Thus, if a node N corresponding to a solution set $W(S)$ is characterized by $S \equiv \{\underline{S}, \overline{S}\}$, then $\underline{S} = \Phi$, i.e., $\underline{S}$ is empty. It will be shown that recognition of a dormant node is instrumental in recognizing translations of solution sets, and that by utilizing this recognition we can obtain higher bounds for the node, as well as fewer branches for the solution tree.

DEFINITION: A *bare* node is a node representing a solution set such that, in tracing back to the start, the first branch encountered corresponds to a prohibition, and there is only one branch corresponding to an allocation.

In this case, the solution-characterizing set $S \equiv \{\underline{S}, \overline{S}\}$ is such that the cardinality of $\underline{S}$ is one. It will be shown that recognition of a bare node is instrumental in recognizing mirror images of solution sets.

First, consider the case when the first partitioning of the solution set is being performed. The solution tree at this stage will look as follows:



If the cycle time of the schedule is m, there exist $m-1$ sets of solutions equivalent to the one represented by node 3 of the above figure. The set of equivalent solution sets are characterized by $\{(1,j)\}$, $\{(2,j)\}$, . . ., $\{(m,j)\}$. Since all these solution sets are equivalent, any one of them can represent all of them. So we must prohibit all except one, e.g. $\{(i,j)\}$. Further, it may be noticed that the set of solutions represented by node 2, i.e., characterized by $\{\overline{(i,j)}\}$, is the set of solutions that includes all the solutions from the sets equivalent to the set characterized by $\{\underline{(i,j)}\}$, except for those which must have element (i,j) in them. But this particular subset of the equivalent solution sets is already included in the solution set represented by node 3. Hence, a way to eliminate further consideration of the sets of solutions equivalent to those represented by node 3 is to eliminate them from node 2. This can be done easily, not only by letting node 2 represent the set of solutions in which item j is prohibited from being scheduled on day i, but also by adding the additional constraint on node 2 that item j must not be scheduled on any day. In this way, consideration of all subsets of solutions equivalent to those represented by node 3 will be eliminated.

Without consideration of translations of solution sets, partitioning was done in the following manner:

Either item j is scheduled on day i,

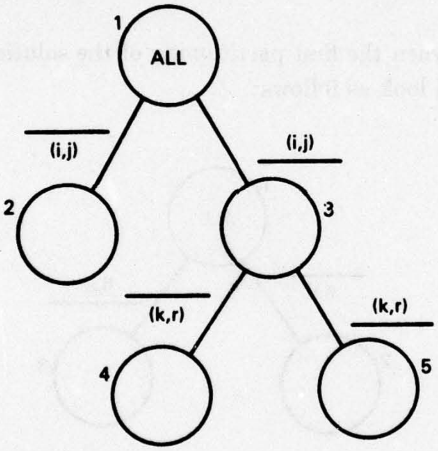
Or item j is barred from being scheduled on day i.

Consideration of translations of solution sets results in a dichotomy of the following kind:

Either item j is scheduled at least once, e.g., on day i,

Or item j is completely eliminated from the schedule.

Elimination of an item j from the schedule will be represented by $(\bar{j})$. So far, we have been talking only of the first dichotomy performed in the branching process. This concept is easily extended to subsequent partitions, provided certain conditions are met. Suppose some node corresponds to prohibitions only, i.e., it is a dormant node, and that each prohibition corresponds to the elimination of the item from the problem, in line with the partitioning scheme mentioned above. Thus, the node corresponds to solution sets in which some items have been taken out of consideration. This is clearly a subproblem of the original problem, with some items eliminated, and hence, one can proceed as if solving just a smaller separation scheduling problem.



Now consider the simplest case of a bare node.

In the above sketch, node 4 is a bare node because by tracing back to node 1 from node 4, the first branch encountered has a prohibition, i.e., $(\bar{k}, \bar{r})$, and the only other branch has an allocation, i.e., $(\underline{i}, \underline{j})$. If node 5 represents the solution set $W(S)$, by the previous discission, an inversion of $W(S)$, e.g., $W(S_1)$, can be constructed by taking the mirror images of all allocations and prohibitions about day i. But $W(S_1)$ is included in the solutions represented by node 4. Hence, as in the case of dormant nodes, a way of prohibiting consideration of $W(S_1)$ is to eliminate it from node 4. This is easily achieved by taking the mirror image of (k, r) about day i, and if this is (g, r), by letting node 4 have the extra prohibition $(\bar{g}, \bar{r})$.

A special case arises when $k=g$. Equivalence cannot be exploited here, as the solution set is its own mirror image.

4. IMPLEMENTATION AND COMPUTATIONAL EXPERIENCE

The algorithms outlined here to solve the SSP and CSSP have been coded in FORTRAN for the CDC6600 computer. A discussion of how the fundamental aspects of the algorithms were implemented follows.

Construction of the Solution Tree

A direct descent [12] approach was employed to execute the partitioning process. Two variables and one array were required to completely define the solution tree and the position on the tree at any stage. One variable identified the current day and the other variable identified the level in the tree of the node currently under consideration. The array contained the indices of those items which received either an allocation or a prohibition on the direct path from the node currently under consideration to the initial node. A negative index indicated a prohibition and a positive index indicated an allocation. Prohibitions or allocations were always made on the days in a sequential manner beginning at day 1, allocations being assigned to the least cost item free on that day. The code continued to make allocations until fathoming took place. After fathoming, the code backtracked up the tree until a node with an allocation was encountered. At that point, the item allocated at that day was prohibited, and the branching process continued down the tree until fathoming again took place.

This type of construction is termed direct descent because the method of branching is predetermined; that is, no search of the tree is required to decide the item and the day to be considered next. The direct descent approach seemed very appropriate here because so few computations were involved in moving from one node in the tree to an adjacent node.

Satisfying the Constraints

An m by n matrix and an n-dimensional vector were utilized to guarantee satisfaction of the separation and upper bound constraints, respectively. The matrix was initialized to zero, and the jth position of the array was assigned the upper bound on the number of times item j could appear during the schedule. When item j was allocated on day i and item h had to be separated from i by d days, a one was added to column h in locations $i \theta d$ through $i \oplus d$. Also, when item j was allocated, a one was subtracted from the jth element of the array. Thus, at any stage, item n could only be allocated on day k if the (k, r)th element of the matrix was zero and the rth element of the array was not zero. When backtracking and a previously allocated item became prohibited, the process was reversed.

Determining Bounds

A lower bound on the best solution to be found in $W(S)$ was obtained by solving a relaxation of the problem conditioned on the assignments made up to that stage. A pass was made from the day currently under consideration onward to day m, which temporarily allocated the least cost item available on each day by considering only the separation constraints incurred as the result of allocations made up to the current day. If the total number of allocations (both temporary and forced by branches above) for an item exceeded the upper bound on the item, the total number of times the item was allocated dropped back to the upper bound, and the difference was added to the number of times that the next least cost item was allocated. A lower bound $L(N)$ for node N was then given by summing the product of the number of times an item was allocated and the item's cost.

The difference between the *SSP* and *CSSP* computer codes *in our implementation* consists of checks to eliminate translations and mirror images of solutions already considered. This yields better bounds, as more prohibitions exist on one side of the tree. With allocations being made sequentially over the days, translations as recognizable through dormant nodes occur only on day 1, and mirror images as recognizable through bare nodes occur only on day 2.

TABLE I. *Computational Results With the SSP and CSSP Algorithms Solving Fifteen Randomly Generated Test Problems*

Problem No.	Problem Size		Time (*CP* seconds)	
	m	*n*	*SSP*	*CSSP*
1	11	30	>360†	22.05
2	7	30	21.70	1.14
3	7	30	68.90	2.66
4	7	30	23.69	1.07
5	14	40	*	98.49
6	8	70	112.10	4.64
7	9	40	36.00	1.87
8	10	30	70.82	5.32
9	21	45	*	191.31
10	7	60	2.90	.20
11	7	60	1.68	.14
12	10	20	50.58	2.49
13	10	20	>300†	10.23
14	14	60	*	242.36
15	21	60	*	232.51

*Indicates problem not attempted with *SSP* algorithm.

†>*N* indicates optimality not verified; run terminated after *N* seconds.

To test the efficiency of the SSP and CSSP computer codes and to compare the two, fifteen sample problems were randomly generated. All sample problems had a separateon matrix approximately 40% dense, and the separation distance varied with the input parameters. Costs were uniformly distributed between n and $n+6n$, 70% of the upper bounds were 1, 25% were 2, and the remainder were 3. The test runs were made on the CDC6600 computer system at the University of Texas at Austin, and all times are CPU seconds required to find the optimal solution and prove optimality. As can be seen, the CSSP algorithum performed remarkably better in every instance; in fact, no attempt was made to solve the four larger problems with the SSP algorithm.

One modification of the algorithm was tested to see what effect better bounds would have on solution times. A transportation problem [3] was solved at each node to obtain a lower bound uniformly stronger than that obtained with the method previously described. Because of the cyclic nature of the separation constraints, the resulting transportation problem was not always trivial. The steps saved in branching did not compensate for the additional time required to solve the transportation problem, and poorer overall times resulted.

5. CONCLUSIONS

This paper has presented a special purpose branch-and-bound algorithm which takes advantage, through the concepts of translation and mirror images, of the cyclic separation scheduling problem's structure. It may also be possible to take a similar advantage when solving other cyclic problems. One trivial extension involves a scheduling problem where an item must be within a certain distance of another item when both are scheduled. The only modification comes in preserving feasibility, where, in our implementation, the locations in the matrix to be incremented or decremented by 1 would change.

In addition to the menu-planning application described in some detail by Balintfy [2], the cyclic separation scheduling model may be useful in an advertising campaign. In this context, several commercials must be scheduled over a specified period to maximize consumer appeal. A modification to the model would arise when certain commercials compliment each other and, if scheduled at all, should be scheduled within a specified time period of one another. This would require the algorithm to combine the techniques discussed in the previous paragraph with those described in Section 4.

BIBLIOGRAPHY

[1] Balintfy, J. L., "Mathematical Models for Menus," Naval Research Reviews, *22* (7), 1–10 (July, 1974).

[2] Balintfy, J. L., "Large-Scale Programming Properties of Menu Planning and Scheduling," in R. W. Cottle and J. Karp (eds.) *Applications of Optimization Methods for Large-Scale Resource-Allocation Problems*, (English Press, 1975).

[3] Charnes, A. and W. W. Cooper, *Management Models and Industrial Applications of Linear Programming*, Vol. I and II, (John Wiley and Sons, Inc., New York, 1961).

[4] Garfinkel, R. S. and G. L. Nemhauser, *Integer Programming*, (John Wiley and Sons, New York, 1972).

[5] Geoffrion, A. and R. E. Marsten, "Integer Programming Algorithms: A Framework and State-of-the-Art Survey," Management Science, *18* (7), 465–491 (May, 1972).

[6] Graves, G. W. and A. B. Whinston, "An Algorithm for the Quadratic Assignment Problem," Management Science, *16* (7), 692–707 (March, 1970).

[7] Lawler, E. L., "The Quadratic Assignment Problem," Management Science, *9* (4), 586–599 (July, 1963).

[8] Little, J. C. D., K. G. Murty, D. W. Sweeney, and C. Karel, "An Algorithm for the Traveling Salesman Problem," Operations Research, *11* (6), 972–989 (November–December, 1963).

AD-A053 613 OFFICE OF NAVAL RESEARCH ARLINGTON VA F/G 15/5
NAVAL RESEARCH LOGISTICS QUARTERLY. VOLUME 24, NUMBER 4.(U)
DEC 77

UNCLASSIFIED NL

2 OF 2
AD
A053613

END
DATE
FILMED
6-78
DDC

[9] Pierce, J. F. and W. B. Crowston, "Tree Search Algorithms in Quadratic Assignment Problems," Naval Research Logistics Quarterly, *18* (1), 1–36 (1971).

[10] Pierskalla, W. P., "The Multidimensional Assignment Problem," Operations Research, *16* (2) 422–431 (March–April, 1968).

[11] Ross, G. T. and R. M. Soland, "A Branch and Bound Algorithm for the Generalized Assignment Problem," Mathematical Programming, *8* (1), 91–103 (February, 1975).

[12] Zoltners, A. A., "A Direct Descent Binary Knapsack Algorithm," Working Paper No. 75–31, School of Business Administration, University of Massachusetts, Amherst (1975).

A PRIMAL SIMPLEX ALGORITHM TO SOLVE A RECTILINEAR DISTANCE FACILITY LOCATION PROBLEM

Ronald D. Armstrong

The University of Texas
Austin, Texas

ABSTRACT

This paper considers the problem of locating m new facilities in the plane so as to minimize a weighted rectangular distance between the new facilities and n existing facilities. A special purpose primal simplex algorithm is developed to solve this problem. The algorithm will maintain at all times a basis of dimension m by m; however, because of the triangularity of the basis matrix, it will not be necessary to form a basis inverse explicitly.

INTRODUCTION

The problem considered in this paper involves assigning positions in the plane for the location of *m* new facilities so as to minimize a weighted rectangular distance between the new facilities and *n* existing facilities. This problem has been studied by Cabot, Francis, and Stary [3], Wesolowsky and Love [15], and Morris [11]. These articles propose a solution by converting the problem to a linear program and then utilizing the simplex algorithm. They also give references to related non-simplex work in the area. Additional references for location theory are given in a review article by Francis and Goldstein [6].

The rectangular distance facility location problem can be stated mathematically as

$$(1)\qquad \text{Minimize} \sum_{i=1}^{m}\sum_{j=m+1}^{n+m} q_{ij}(|x_{1i}-a_{1j}|+|x_{2i}-a_{2j}|)+\sum_{i=1}^{m-1}\sum_{j=i+1}^{m} q_{ij}(|x_{1i}-x_{1j}|+|x_{2i}-x_{2j}|),$$

where (x_{1i}, x_{2i}) is the location of the *i*th new facility in the plane, (a_{1j}, a_{2j}) is the location of the *j*th existing facility in the plane, and q_{ij} is a positive weight on the rectilinear distance between facilities *i* and *j*.

A more general statement of the problem of finding parameters to minimize the rectangular distances between given points and associated linear functions of these parameters is given by

$$(2)\qquad \text{Minimize} \sum_{j=1}^{s} |d_j{}^T x - b_j|,$$

where the scalar b_j and the t-dimensional vector d_j are given for all j, and x is the vector of parameters which must be determined. Here, the positive weights, if present, are brought inside the absolute value sign and are reflected in b_j and d_j. Two common names given to (2) are least absolute deviations curve-fitting problems and L_1 norm regression problems.

Applications which give rise to problem (2) are found in statistics [2], numerical analysis [12], and management science [4, 5, 10]. Charnes, Cooper, and Ferguson [5] were the first to utilize linear programming to solve (2), and it is generally recognized that this is the most efficient method to solve the problem [1, 13]. A linear programming equivalent of (2) is

$$\text{Minimize} \sum_{j=1}^{s} (P_j+N_j), \tag{3}$$

subject to

$$b_j-d_j^T x-P_j+N_j=0,$$

$$P_j\geq 0 \text{ and } N_j\geq 0,\ j=1, 2, \ldots, s.$$

Wagner [14] demonstrates how, upon taking the dual of (3), the problem becomes one with upper bound restrictions on all the dual variables and, hence, would require the working basis to be a t by t matrix. This makes the problem much more tractable with existing linear programming codes, because, generally, t will be substantially less than s. It is an adaptation of Wagner's approach that is used by Cabot, Francis, and Stary [3] in reducing (1) to a network problem.

Because of the special structure of (3), it can also be solved with a primal algorithm employing a working basis of size t by t. Barrodale and Roberts [1] and Spyropoulos, Kiountouzis, and Young [13] report superior computational results in solving the primal problem directly with a special purpose algorithm. An extension of such an algorithm will be presented here to solve (1).

A LINEAR PROGRAMMING FORMULATION

An initial observation concerning (1) is that it decomposes immediately into two disjoint problems—one for each coordinate in the plane. Because of this, we will now drop the double subscript notation on X and a, and refer to the problem.

$$\text{Minimize} \sum_{i=1}^{m} \sum_{j=m+1}^{n+m} q_{ij}|X_i-a_j|+\sum_{i=1}^{m-1} \sum_{j=i+1}^{m} q_{ij}|X_i-X_j|. \tag{4}$$

Cabot, Francis, and Stary [3] show that (4) is equivalent to

$$\text{Minimize} \sum_{i=1}^{m} \sum_{j=i+1}^{n+m} q_{ij}(P_{ij}+N_{ij}), \tag{5}$$

subject to

$$X_i+P_{ij}-N_{ij}=a_j,\ i=1, 2, \ldots, m$$

$$j=m+1,\ m+2, \ldots, n+m;$$

$$X_i-X_j+P_{ij}-N_{ij}=0,\ i=1, 2, \ldots m-1;$$

$$j=i+1, \ldots, m;$$

$$N_{ij}\geq 0 \text{ and } P_{ij}\geq 0, \text{ for all } (i, j),$$

where N_{ij} and P_{ij} are, respectively, the negative and positive absolute deviation of facility i from facility j in the coordinate under consideration.

It can be shown [3] that the dual of (5) is a capacitated network problem, once additional variables and constraints have been added. A primal algorithm will be presented here to solve a slightly modified version of (5). Although all the results given here can be obtained by considering only (5), it is felt that the following equivalent problem makes the development more intuitive.

(6) $$\text{Minimize} \sum_{i=1}^{m}\sum_{j=i+1}^{n+m} q_{ij}(P_{ij}+N_{ij}),$$

subject to

$$a_j - P_{ij} \le X_i \le a_j + N_{ij},\quad i=1, 2, \ldots, m;$$

$$j=m+1,\ m+2, \ldots, m+n;$$

$$-P_{ij} \le X_i - X_j \le N_{ij},\quad i=1, 2, \ldots, m-1;$$

$$j=i+1, \ldots, m;$$

$$N_{ij} \ge 0 \text{ and } P_{ij} \ge 0 \text{ for every } (i, j).$$

In matrix notation, the constraints of (6) can be written

(7) $$\binom{a}{0} - P \le Ax \le \binom{a}{0} + N,$$

$$N \ge 0 \text{ and } P \ge 0.$$

It is easily shown [1, 13] from fundamental linear programming theory that an optimal basis for (6) (or (5)) will have all mx_j's in the basis, and at least m constraints will have $P_{ij}=N_{ij}=0$. Furthermore, the submatrix of A formed from these m constraints will be linearly independent and, once they have been chosen, the remaining basic variables can be determined. To illustrate this, we partition (7) in the following manner:

$$\begin{bmatrix} F_B \\ F_R \end{bmatrix} - \begin{bmatrix} P_B \\ P_R \end{bmatrix} \le \begin{bmatrix} B \\ R \end{bmatrix} X \le \begin{bmatrix} F_B \\ F_R \end{bmatrix} + \begin{bmatrix} N_B \\ N_R \end{bmatrix},$$

where B is an m by m nonsingular matrix corresponding to the m constraints previously mentioned. We now make the transformation $z=BX$, and define a new equivalent problem:

(8) $$\text{Minimize} \sum_{i=1}^{m}\sum_{j=i+1}^{n+m} q_{ij}(P_{ij}+N_{ij}),$$

subject to

$$\begin{bmatrix} F_B \\ F_R \end{bmatrix} - \begin{bmatrix} P_B \\ P_R \end{bmatrix} \le \begin{bmatrix} I \\ RB^{-1} \end{bmatrix} z \le \begin{bmatrix} F_B \\ F_R \end{bmatrix} + \begin{bmatrix} N_B \\ N_R \end{bmatrix},$$

$$N \ge 0 \text{ and } P \ge 0.$$

Since $P_B=N_B=0$, the current "basic" solution has $\bar{z}=F_B$, $\bar{X}=B^{-1}\bar{z}$; $\bar{P}_{ij}$ and $\bar{N}_{ij}$ are determined from the constraints while making the deviations as small as possible.

We will now demonstrate how several standard simplex pivots may be combined into one pivot. We initially choose a B which is our working basis and define two index sets

$$I^{+}=\{(i,j)\,|\,P_{ij}>0,\ (i,j)\in R\}$$

and

$$I^{-}=\{(i,j)\,|\,N_{ij}>0,\ (i,j)\in R\},$$

where R is the index set of P_{ij} in P_R (or N_{ij} in N_R). In the case of degeneracy ($P_{ij}=N_{ij}=0$), (i, j) may be assigned initially to either I^- or I^+. The objective value associated with this basis is

$$\sum_{(i,j)\in I^+} q_{ij}\overline{P}_{ij}+\sum_{(i,j)\in I^-} q_{ij}\overline{N}_{ij}=\sum_{(i,j)\in I^+} q_{ij}(F_{ij}-R_{ij}B^{-1}\overline{z})+\sum_{(i,j)\in I^-} q_{ij}(R_{ij}B^{-1}\overline{z}-F_{ij},$$

where R_{ij} is a *row* of R and F_{ij} is an *element* of F; that is, $F_{ij}=0$ or a_j. It is now seen that by varying an element of z, e.g. z_k, away from its current value $\overline{z}_k$ by δ, and leaving the remaining elements of z fixed, the objective value will change by

$$\sum_{(i,j)\in I^-} q_{ij}R_{ij}B_k^{-1}\delta-\sum_{(i,j)\in I^+} q_{ij}R_{ij}B_k^{-1}\delta+|\delta q_{\alpha\beta}^{(k)}|=\delta\Delta_k+|\delta q_{\alpha\beta}^{(k)}| \tag{9}$$

where B_k^{-1} is the kth *column* of B^{-1} and $q_{\alpha\beta}^{(k)}$ is the weight associated with the kth row in the working basis. In other words, if $|\Delta_k|$ is greater than $|q_{\alpha\beta}|$, then the objective can be decreased by varying z_k in the direction prescribed by the sign of Δ_k ($\Delta_k<0$ indicates an increase, and $\Delta_k>0$ indicates a decrease).

The amount that z_k can vary away from $\overline{z}_k$ before the rate of change in the objective becomes nonegative is given by ordering the following ratios:

$$\rho_{ij}=\begin{cases}\dfrac{\overline{P}_{ij}}{-\operatorname{sgn}(\Delta_k)R_{ij}B_k^{-1}};\ \operatorname{sgn}(\Delta_k)R_{ij}B_k^{-1}<0,\ (i,j)\in I^+\\[2ex] \dfrac{\overline{N}_{ij}}{\operatorname{sgn}(\Delta_k)R_{ij}B_k^{-1}};\ \operatorname{sgn}(\Delta_k)R_{ij}B_k^{-1}>0,\ (i,j)\in I^-\end{cases}$$

Let $L(u)$ denote the index set of the rows yielding the u smallest ratios. After $|\delta|$ equals the minimum ratio (ρ_{ij}, $(i, j)\in L(1)$), the indices (i, j) will move from I^+ to I^- (or from I^- to I^+) and the objective change given by (9) will be updated accordingly, However, the change will be nonnegative until the (i, j) present in $L(h)$ but not in $L(h-1)$ moves from one index set to another, where

$$|\Delta_k|-\sum_{(i,j)\in L(h-1)} 2|q_{ij}R_{ij}B_k^{-1}|\geq q_{\alpha\beta}^{(k)},$$

$$|\Delta_k|-\sum_{(i,j)\in L(h)} 2|q_{ij}R_{ij}B_k^{-1}|\leq q_{\alpha\beta}^{(k)}.$$

The (i, j) row present in $L(h)$ but not in $L(h-1)$ will not enter the working basis in place of (α, β). This iteration of the algorithm corrcsponds to h standard simplex iterations.

The computational simplifications which are possible because of the special structure of the A matrix will be discussed in the next section.

IMPLEMENTATION FOR FACILITY LOCATION AND A SAMPLE PROBLEM

If the problem to be solved is of the form given by (6), an efficient implementation of the algorithm outlined in the previous section requires an adaptation of the techniques utilized in solving network problems. Hence, the basis matrix B will never be inverted, but rather saved in a spanning tree format [7, 8], and the triangularity of B will allow a direct solution of linear systems involving B. In order to facilitate the relationship between (6) and a standard capacitiated network problem, it is conceptually convenient to add a dummy variable X_0 to the first (m) (n) constraints. These constraints would then appear as

$$a_j - P_{ij} \leq X_i - X_0 \leq a_j + N_{ij},\ i=1, 2, \ldots, m,\ j=m+1, m+2, \ldots, m+n,$$

with the additional restriction that $X_0=0$. The transpose of the augmented A matrix now has a network structure.

The linear system $B\ X=F_B$ can be solved efficiently through the use of link list structures [7, 8] in the same manner as the dual variables are obtained in the standard network problem. That is, X_0 (which will always be at the root of the tree) is set equal to zero, and the remaining variables are obtained by a single pass through the tree. Because the sweep of the tree begins by setting $X_0=0$, the necessity of self-loop (slack) arcs is alleviated and the spanning tree structure is preserved.

To check for the optimality of the current solution, and to determine the row to leave the basis, the system

$$\Delta B = (\Delta_1, \Delta_2, \ldots, \Delta_m)\ B = \theta$$

must be solved for Δ, where

$$\theta = \sum_{(i,j)\in I^-} q_{ij} R_{ij} - \sum_{(i,j)\in I^+} q_{ij} R_{ij}.$$

The values for Δ can be obtained by working back to the root of the tree, and the additional equation which may be created by the dummy column is redundant.

To demonstrate the algorithm, a sample problem with three existing facilities and three new facilities will be solved in one coordinate. The locations of the existing facilities are given by $a_1=7$, $a_2=11$, and $a_3=16$. The matrix of weights, or q_{ij}'s, is given in Table 1.

TABLE 1. *Weights on the Rectilinear Distances Between Facilities in the Sample Problem*

		Facility					
		1	2	3	4	5	6
Facility	1	---	2	1	1	1	6
	2	2	---	1	1	1	3
	3	1	1	---	8	1	1

The initial basis {(1, 5), (2, 5), (3, 5)} places the three new facilities at the location of facility 5. The spanning tree representation of this basis and the basis at other iterations is given in Figure 1.

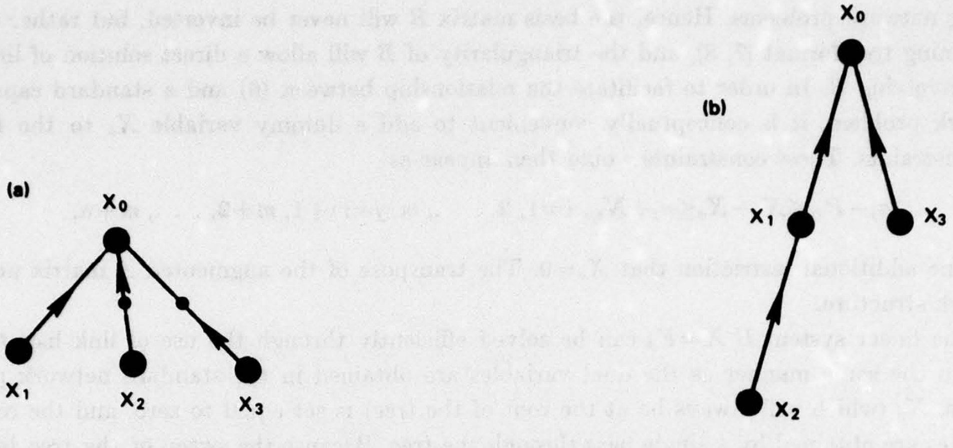


FIGURE 1. (a) Spanning-tree representation of the first three bases. (b) Spanning-tree representation of the final basis. Although the first three bases have the same representation, the node potentials of the tree are different.

With the first basis, we have

$I^+=\{(1, 6), (2, 6), (3, 6), (1, 2), (1, 3), (2, 3)\}$

$I^-=\{(1, 4), (2, 4), (3, 4)\}$

and $\Delta=(\Delta_1, \Delta_2, \Delta_3)=(-8, -1, 9)$. Row (3, 5) is chosen to leave the basis and row (3, 4) enters the basis. In this situation, row (3, 4) yields the third smallest ratio; hence, the first iteration of the algorithm combines three standard simplex iterations into one.

The second basis is $\{(1, 5), (2, 5), (3, 4)\}$ and has

$I^+=\{(1, 6), (2, 6), (1, 2), (3, 5), (3, 6)\}$

$I^-=\{(1, 4), (2, 4), (1, 3), (2, 3)\}$

and $\Delta=(-6, 1, -5)$. Row (1, 5) leaves the basis and (1, 6) enters the basis. Row (1, 6) yields the second smallest ratio; hence, this iteration combines two standard simplex iterations into one.

The third basis is $\{(1, 6), (2, 5), (3, 4)\}$ and has

$I^+=\{(2, 6), (3, 5), (3, 6)\}$

$I^-=\{(1, 4), (1, 5), (2, 4), (1, 3), (2, 3), (1, 2)\}$

and $\Delta=(5, -3, -4)$. Row (2, 5) leaves the basis and row (1, 2) enters the basis.

The fourth basis is $\{(1, 6), (1, 2), (3, 4)\}$ and has

$I^+=\{(2, 6), (3, 5), (3, 6)\}$

$I^-=\{(1, 4), (1, 5), (2, 4), (2, 5), (1, 3), (2, 3)\}$

and $\Delta=(3, 0, -4)$. This basis is optimal because $q_{16}=6$, $q_{12}=2$ and $q_{34}=8$. The optimal objective value is 59 with $X_1=16$, $X_2=16$ and $X_3=7$.

A detailed description of the computational steps required, such as updating the spanning tree, will not be given here, but they are analogous to those found in capacitated network algorithms. In particular, an efficient procedure for calculating the ratios, ρ_{ij}, parallels the method for calculating the ratios in a dual simplex network code [9]. This is what one would expect, as the dual of (5) is a capacitated network problem, and solving the duel with a dual simplex algorithm is the same as solving the primal with a primal simplex algorithm. The advantages to our algorithm are: (a) a feasible solution to (b) is immediately obtained by choosing any spanning tree and (b) several simplex iterations may be combined into one.

REFERENCES

[1] Barrodale, I., and F. D. K. Roberts, "An Improved Algorithm for Discrete L_1 Linear Approximation," SIAM Journal of Numerical Analysis, *10*, 839–848 (1973).

[2] Barrodale, I., "L_1 Approximation and the Analysis of Data," Applied Statistics, *17*, 51–57 (1968).

[3] Cabot, A. V., R. L. Francis, and M. A. Stary, "A Network Flow Solution to a Rectilinear Distance Facility Location Problem," AIIE Transactions, *2*, 132–141 (1970).

[4] Charnes, A., and W. W. Cooper, *Management Models and Industrial Applications*, Vol. I and II (John Wiley & Sons, Inc., New York, 1961).

[5] Charnes, A., W. W. Cooper, and R. Ferguson, "Optimal Estimation of Executive Compensation by Linear Programming," Management Science, *2*, 138–151 (1955).

[6] Frances, R. L. and J. M. Goldstein, "Location Theory: A Selective Bibliography," Operations Research, *22*, 400–410 (1974).

[7] Glover, F., D. Karney, and D. Klingman, "The Augmented Predecessor Index Method for Locating Stepping Paths and Assigning Dual Prices in Distribution Problems," Transportation Science, *6*, 171–180 (1972).

[8] Glover, F., D. Karney, D. Klingman, and A. Napier, "A Computational Study on Start Procedures, Basis Change Criteria, and Solution Algorithms for Transportation Problems," Management Science, *20*, 793–814 (1974).

[9] Glover, F., D. Klingman, and A. Napier, "An Efficient Dual Approach to Network Problems" Opsearch *9*(1), 1–19 (1972).

[10] Lee, S. M., *Goal Programming for Decision Analysis*, (Auerbach, Philadelphia, 1972).

[11] Morris, J. G., "A Linear Programming Solution to the Generalized Rectangular Distance Weber Problem," Naval Research Logistics Quarterly, *22*, 155–164 (1975).

[12] Rabinowitz, P., "Applications of Linear Programming to Numerical Analysis," SIAM Review, *10*, 121–159 (1968).

[13] Spyropoulos, K., E. Kiountouzis, and A. Young, "Discrete Approximation in the L_1 Norm," The Computer Journal, *16*, 180–186 (1973).

[14] Wagner, H. M., "Linear Programming Techniques for Regression Analysis," Journal of the American Statistical Association, *54*, 206–212 (1959).

[15] Wesolowsky, G. O. and R. F. Love, "The Optimal Location of New Facilities Using Rectangular Distances," Operations Research, *19*, 124–130 (1971).

MONOTONE FAILURE RATES FOR MULTIVARIATE DISTRIBUTIONS*

Henry W. Block

University of Pittsburgh
Pittsburgh, Pennsylvania

ABSTRACT

It is shown that the monotone multivariate failure rates of Brindley and Thompson have no natural analog involving the multivariate failure rate function of Basu for absolutely continuous distributions. Quantities related to the multivariate failure rate function are used to define monotone failure rates. It is shown that these are equivalent to the monotone failure rates of Brindley and Thompson. Based on these quantities, the loss of memory property of Marshall and Olkin is characterized.

1. INTRODUCTION

For multivariate distributions, many different concepts of monotone failure rate can be defined. Several concepts of monotone failure rates were introduced by Brindley and Thompson [5]. For absolutely continuous bivariate random variables, Basu [2] introduced the multivariate failure rate function. As in the univariate case, it seems reasonable to attempt to define increasing and decreasing failure rates in terms of this multivariate failure rate function. In Section 3, we attempt to show that if this is done using this function, the resulting concepts do not parallel the monotone failure rates of Brindley and Thompson.

In Section 4, we utilize functions related to the multivariate failure rate function of Basu in order to define monotone failure rates. These functions turn out to be the components of a hazard gradient introduced independently by Johnson and Kotz [7] and Marshall [8]. The uses of these functions by these authors and by the present author are quite different. Connections will be discussed in a subsequent paper. It is shown that these monotone failure and hazard rates correspond to the monotone failure rates of Brindley and Thompson. Some consequences are given. Based on the concepts of Section 4, a characterization of the loss of memory property (LMP) of Marshall and Olkin [9] is given in Section 5. A concept of increasing hazard rate due to Harris [6] which is less important than the concept of Brindley and Thompson is discussed in the appendix.

*Partially supported by the Office of Naval Research under Contract N00014-75-C-0781 with the University of California and Contract N00014-76-C-0839 with the University of Pittsburgh.

2. FAILURE RATES

In discussing monotone failure rates, the terms "increasing" and "decreasing" will mean, respectively, nondecreasing and nonincreasing. Since we mainly have in mind survival times, when we discuss n random variables $X_1, \ldots, X_n$, we shall assume that $P(X_1>0, \ldots, X_n>0)=1$ and that the arguments of all distribution and density functions are positive. An abuse of terminology which we use is calling $\overline{F}(x_1, \ldots, x_n)=P(X_1>x_1, \ldots, X_n>x_n)$ a distribution function (d.f.).

In the univariate case, the failure rate function

$$r(x)=\frac{f(x)}{\overline{F}(x)}$$

for absolutely continuous distributions is increasing (decreasing) if and only if

$$\frac{F(x+\Delta)-F(x)}{\overline{F}(x)}=1-\frac{\overline{F}(x+\Delta)}{\overline{F}(x)}$$

is increasing (decreasing) in x for each Δ, i.e., $\overline{F}(x+\Delta)/\overline{F}(x)$ is decreasing (increasing) in x for each Δ. In either case, we say that the distribution is IFR (DFR).

This second concept has been generalized by Harris [6] and by Brindley and Thompson [5], and both of these generalizations involve the assumption that

$$\frac{\overline{F}(x_1+\Delta, \ldots, x_n+\Delta)}{\overline{F}(x_1, \ldots, x_n)} \tag{2.1}$$

is decreasing (increasing) in $x_1, \ldots, x_n$ for each Δ. We give the definition of Brindley and Thompson below. The concept of Harris, which is of less interest, will be discussed in the appendix.

DEFINITION: A multivariate d.f. $\overline{F}(x_1, \ldots, x_n)$ for the random variables $\mathbf{X}_1, \ldots, \mathbf{X}_n$ has *increasing failure rate* (IFR) if (*) below is satisfied, and has *joint* IFR if, in addition, (*) is satisfied for all subfamilies of the n random variables. This latter concept is designated MIFR by Barlow and Proschan [1].

$$(*)\, P(X_1>x_1+\Delta, \ldots, X_n>x_n+\Delta \mid X_1>x_1, \ldots, X_n>x_n)=\frac{\overline{F}(x_1+\Delta, \ldots, x_n+\Delta)}{\overline{F}(x_1, \ldots, x_n)} \text{ is decreasing in } x_1, \ldots,$$
x_n for each Δ.

Decreasing failure rate (DFR) and *joint* DFR are defined by interchanging the word increasing for decreasing in (*).

For absolutely continuous random variables, the failure rate function has been generalized to the multivariate case by Basu [2] and is given in the following definition.

DEFINITION: Let $\overline{F}(x_1, \ldots, x_n)$ have an absolutely continuous distribution with density $f(x_1, \ldots, x_n)$. Then

$$r(x_1, \ldots, x_n)=\frac{f(x_1, \ldots, x_n)}{\overline{F}(x_1, \ldots, x_n)} \tag{2.2}$$

is called the *multivariate failure rate function*.

To our knowledge, the concept of monotone failure rate for this function has not been investigated.

A concept related to the monotone failure rates above is the loss of memory property (LMP) of Marshall and Olkin [9]. A distribution is said to have the LMP if

$$P(X_1>x_1+\Delta, \ldots, X_n>x_n+\Delta | X_1>x_1, \ldots, X_n>x_n)=P(X_1>\Delta, \ldots, X_n>\Delta)$$

or, equivalently,

$$\bar{F}(x_1+\Delta, \ldots, x_n+\Delta)=\bar{F}(x_1, \ldots, x_n)\bar{F}(\Delta, \ldots, \Delta) \tag{2.3}$$

for all $x_1, \ldots, x_n$ and Δ. This property is possessed by the multivariate exponential distribution (MVE) of Marshall and Olkin [9], the bivariate exponential extension of Block and Basu [4], and the multivariate exponential extension of Block [3]. It characterizes these distributions when the marginals are specified.

3. THE BIVARIATE FAILURE RATE FUNCTION

As mentioned in the previous section, we are aware of no study of monotone failure rate for the multivariate failure rate function (2.2). One possible reason for this is that there do not appear to be close analogies between monotone failure rates for this function and the monotone failure rate concepts of Brindley and Thompson [5]. In the present section, we demonstrate this lack of close relationship in the bivariate situation.

In attempting to define monotone failure rates using the bivariate failure rate function, two possibilities seem most natural. The first of these represents a straightforward mathematical generalization from the univariate case.

CONDITION 3.1:

$$r(x_1, x_2)=\frac{f(x_1, x_2)}{\bar{F}(x_1, x_2)}$$

is increasing (decreasing) in x_1 and x_2.

This condition, however, appears too strong, since the boundary between the increasing and the decreasing cases is that $f(x_1, x_2)/\bar{F}(x_1, x_2)$ is constant. In the case of exponential marginals, Basu [2] showed that this implies that the bivariate distribution has independent marginals. Puri [10] showed that this was also true under the apparently weaker assumption that at least one of the marginals is exponential. If the marginals are unspecified, Puri and Rubin [11] showed that the joint distribution must be a mixture of independent exponentials; therefore, the marginals are mixtures of univariate exponentials. These in general have a decreasing failure rate, which is not very satisfactory for a boundary distribution.

A second condition, implied by the first condition, is based on the observation that for the bivariate exponential extension of Block and Basu [4], the bivariate failure rate function $r(x_1+\Delta, x_2+\Delta)$ is constant in Δ for each x_1 and x_2. As shown in Section 5, this property is shared by any absolutely continuous distribution having the LMP.

CONDITION 3.2:

$$r(x_1+\Delta, x_2+\Delta)$$

is increasing (decreasing) in Δ for all x_1 and x_2.

The following examples will show that these concepts are not closely related to the monotone failure rates discussed in the previous section, although the second concept is related to the LMP, as will be shown in Theorem 5.2.

Condition 3.1 implies neitner the IFR (DFR) nor the joint IFR (DFR). This is shown in Example 1. This example also shows that Condition 3.2 implies neither of these properties, since it is weaker than the first property. The second example shows that neither the IFR (DFR) nor the joint IFR (DFR) implies the increasing version of Condition 3.1.

EXAMPLE 1: Let

$$\bar{F}(x_1, x_2)=p \exp(-(\lambda_1 x_1+\lambda\cdot\lambda_1^{-1} x_2))+(1-p) \exp(-(\lambda_2 x_1+\lambda\cdot\lambda_2^{-1} x_2))$$

where $0<p<1$ and $\lambda, \lambda_1, \lambda_2>0$. Then

$$\frac{f(x_1, x_2)}{\bar{F}(x_1, x_2)}=\lambda,$$

so Conditions 3.1 and 3.2 are clearly satisfied. However, it can be shown that, for certain choices of λ_1, λ_2, and λ,

$$\frac{\bar{F}(x_1+t, x_2+t)}{\bar{F}(x_1, x_2)}$$

is strictly decreasing in x_1 and strictly increasing in x_2, so the distribution is neither *IFR* nor *DFR*.

EXAMPLE 2: Let

$$\bar{F}(x_1, x_2)=\frac{\lambda}{\lambda_1+\lambda_2} \exp(-\lambda_1 x_1-\lambda_2 x_2-\lambda_{12} \max(x_1, x_2))-\frac{\lambda_{12}}{\lambda_1+\lambda_2} \exp(-\lambda \max(x_1, x_2))$$

where $\lambda_1, \lambda_2, \lambda_{12}>0$ and $\lambda=\lambda_1+\lambda_2+\lambda_{12}$, which is the bivariate exponential extension of Block and Basu [4]. This distribution has the *LMP* which implies that it is *IFR*, and it can be shown that the marginals are *IFR*. Thus, it is joint *MIFR*. However,

$$r(x_1, x_2)=\begin{cases} \dfrac{\lambda_1(\lambda_2+\lambda_{12})\lambda}{\lambda-\lambda_{12}\exp(-\lambda_1(x_2-x_1))} \text{ if } x_1<x_2 \\ \dfrac{(\lambda_1+\lambda_{12})\lambda_2\lambda}{\lambda-\lambda_{12}\exp(-\lambda_2(x_1-x_2))} \text{ if } x_2<x_1, \end{cases}$$

which is strictly increasing in x_1 for $x_1<x_2$ but strictly decreasing in x_1 for $x_1>x_2$, so that the first condition is not satisfied.

Since the bivariate failure rate function does not seem to be closely related to the monotone failure rates discussed previously, in the next section we introduce functions which lead to failure rates which parallel those formerly discussed.

4. ALTERNATE MULTIVARIATE FAILURE RATES

Since the bivariate failure rate function

$$r(x_1, x_2)=\frac{f(x_1, x_2)}{\bar{F}(x_1, x_2)}=\frac{\bar{F}_{x_1, x_2}(x_1, x_2)}{\bar{F}(x_1, x_2)},$$

where

$$\overline{F}_{x_1, x_2}(x_1, x_2)=\frac{\partial^2}{\partial x_1 \partial x_2}\overline{F}(x_1, x_2),$$

does not appear to be compatible with the monotone failure rates of Brindley and Thompson [5], we consider the functions

$$\frac{\overline{F}_{x_1}(x_1, x_2)}{\overline{F}(x_1, x_2)} \text{ and } \frac{\overline{F}_{x_2}(x_1, x_2)}{\overline{F}(x_1, x_2)}$$

in the bivariate case, or more generally

$$\frac{\frac{\partial}{\partial x_i}\overline{F}(x_1, \ldots, x_n)}{\overline{F}(x_1, \ldots, x_n)} \text{ for } i=1,2, \ldots, n$$

in the multivariate case. The negatives of these functions are the components of the hazard gradient mentioned in the introduction, and give the ith one-dimensional failure rate functions upon setting $x_j=0$ for $j\neq i$.

DEFINITION: The d.f. $\overline{F}(x_1, \ldots, x_n)$ for the random variables $X_1, \ldots, X_n$ is IFR_a (DFR_a) if (**) below is satisfied, and is *joint* IFR_a (*joint* DFR_a) if, in addition, (**) is satisfied for all subfamilies of the n random variables. It is assumed in the following that the first partial derivatives exist.

(**) For each $i=1,2, \ldots, n$,

$$\frac{\overline{F}_{x_i}(x_1+\Delta, \ldots, x_n+\Delta)}{\overline{F}(x_1+\Delta, \ldots, x_n+\Delta)}=\frac{\frac{\partial}{\partial x_i}\overline{F}(x_1+\Delta, \ldots, x_n+\Delta)}{\overline{F}(x_1+\Delta, \ldots, x_n+\Delta)}$$

is decreasing (increasing) in Δ for each $x_1, \ldots, x_n$.

The following properties hold for random variables $X_1, \ldots, X_n$ with d.f. $\overline{F}(x_1, \ldots, x_n)$ which are joint IFR_a (joint DFR_a):

(1) A single random variable which is joint IFR_a (joint DFR_a) is IFR (DFR) in the univariate sense.

(2) The union of two mutually independent sets of joint IFR_a (joint DFR_a) random variables is joint IFR_a (joint DFR_a).

(3) Any subset of joint IFR_a (joint DFR_a) random variables is joint IFR_a (joint DFR_a).

(4) Sets of minimums of joint IFR_a (joint DFR_a) random variables are joint IFR_a (joint DFR_a).

The proofs of (1) and (3) are by definition and (2) is easy to show.

PROOF OF (4):

Let $Y_1, \ldots, Y_m$ be joint IFR_a (joint DFR_a) and $X_i=\min_{j\in J_i} Y_j$ for $i=1,2, \ldots, n$ where $J_i \subset\{1, 2, \ldots, m\}$. Let

$$I_j=\left\{i \middle| j\in J_i\right\},\ y_j=\max_{i\in Ij} x_i, \text{ and } K_k=\left\{y_l \middle| y_l=x_k\right\}.$$

Then

$$\frac{\frac{\partial}{\partial x_k} P\left\{\bigcap_{i=1}^{n} \{X_i > x_i + \Delta\}\right\}}{P\left\{\bigcap_{i=1}^{n} \{X_i > x_i + \Delta\}\right\}} = \frac{\frac{\partial}{\partial x_k} P\left\{\bigcap_{j=1}^{m} \{Y_j > y_j + \Delta\}\right\}}{P\left\{\bigcap_{j=1}^{m} \{Y_j > y_j + \Delta\}\right\}} = \frac{\frac{\partial}{\partial x_k} P\left\{\bigcap_{j\epsilon K_k} \{Y_j > y_j + \Delta\} \cap \bigcap_{j\epsilon K_k} \{Y_j > x_k + \Delta\}\right\}}{P\left\{\bigcap_{j=1}^{m} \{Y_j > y_j + \Delta\}\right\}}$$

$$= \frac{\sum_{l\epsilon K_j}\left[\frac{\partial}{\partial x_k} P\left\{\bigcap_{j\epsilon K_k} \{Y_j > y_j + \Delta\} \cap \bigcap_{\substack{j\epsilon K_k\\ j\neq l}} \{Y_j > y_j + \Delta\} \cap \{Y_l > x_k + \Delta\}\right\}\right]}{P\left\{\bigcap_{j=1}^{m} \{Y_j > y_j + \Delta\}\right\}},$$

where the quantity in brackets is evaluated at $y_j = x_k$ for $j\epsilon K_k$ and each of the summands is decreasing (increasing) in Δ, since $Y_1, \ldots, Y_m$ are joint IFR_a (joint DFR_a). This implies that $X_1, \ldots, X_n$ are joint IFR_a (joint DFR_a).

It is not surprising that random variables which are joint IFR_a (joint DFR_a) satisfy the preceding four conditions (see the six conditions of Section 2 of Brindley and Thompson [5]) in the light of the following proposition.

THEOREM 4.1: Subject to the existence of first partial derivatives of the *d.f.* $\bar{F}$, $\bar{F}$ is IFR_a (DFR_a) if and only if $\bar{F}$ is IFR (DFR).

PROOF: Assume $\bar{F}$ is IFR_a (DFR_a). Then for each $i = 1, \ldots, n$,

$$\frac{\bar{F}_{x_i}(x_1+\Delta, \ldots, x_n+\Delta)}{\bar{F}(x_1+\Delta, \ldots, x_n+\Delta)}$$

is decreasing (increasing) in Δ for fixed $x_1, \ldots, x_n$. Thus, for each $i = 1, \ldots, n$ and $\Delta > 0$,

$$\frac{\bar{F}_{x_i}(x_1, \ldots, x_n)}{\bar{F}(x_1, \ldots, x_n)} \geq (\leq) \frac{\bar{F}_{x_i}(x_1+\Delta, \ldots, x_n+\Delta)}{\bar{F}(x_1+\Delta, \ldots, x_n+\Delta)}.$$

It then follows for each $i = 1, \ldots, n$ and $\Delta > 0$ that

$$0 \geq (\leq) \frac{\bar{F}(x_1, \ldots, x_n)\bar{F}_{x_i}(x_1+\Delta, \ldots, x_n+\Delta) - \bar{F}_{x_i}(x_1, \ldots, x_n)\bar{F}(x_1+\Delta, \ldots, x_n+\Delta)}{(\bar{F}(x_1, \ldots, x_n))^2}$$

$$= \frac{\partial}{\partial x_i} \frac{\bar{F}(x_1+\Delta, \ldots, x_n+\Delta)}{\bar{F}(x_1, \ldots, x_n)}.$$

Thus,

$$\frac{\bar{F}(x_1+\Delta, \ldots, x_n+\Delta)}{\bar{F}(x_1, \ldots, x_n)}$$

is decreasing (increasing) in $x_1, \ldots, x_n$ for each Δ, and thus is IFR (DFR). Reversing the argument gives the proof in the opposite direction.

COROLLARY: Under the assumptions of Theorem 4.1 on $\bar{F}$, $\bar{F}$ is joint IFR_a (joint DFR_a) if and only if $\bar{F}$ is joint IFR (joint DFR).

5. A CHARACTERIZATION OF THE LOSS OF MEMORY PROPERTY

We now turn to the LMP and obtain a characterization which is related to the preceding concepts. Upon differentiating (2.3) repeatedly and then substituting (2.3) into the resulting equation, we obtain, for each positive integer m and positive integers $k_1, \ldots, k_n$ such that $k_1+k_2+\ldots+k_n=m$

$$\frac{\frac{\partial^m}{\partial x_1^{k_1}\partial x_2^{k_2}\cdots\partial x_n^{k_n}}\bar{F}(x_1+\Delta, \ldots, x_n+\Delta)}{\bar{F}(x_1+\Delta, \ldots, x_n+\Delta)}=\frac{\frac{\partial^m}{\partial x_1^{k_1}\partial x_2^{k_2}\cdots\partial x_n^{k_n}}\bar{F}(x_1, \ldots, x_n)}{\bar{F}(x_1, \ldots, x_n)}$$

which we can express by saying that

$$\frac{\frac{\partial^m}{\partial x_1^{k_1}\partial x_2^{k_2}\cdots\partial x_n^{k_n}}\bar{F}(x_1, \ldots, x_n)}{\bar{F}(x_1, \ldots, x_n)}$$

is stationary in $x_1, \ldots, x_n$ or that

$$\frac{\frac{\partial^m}{\partial x_1^{k_1}\partial x_2^{k_2}\cdots\partial x_n^{k_n}}\bar{F}(x_1+\Delta, \ldots, x_n+\Delta)}{\bar{F}(x_1+\Delta, \ldots, x_n+\Delta)}$$

is constant in Δ for all $x_1, \ldots, x_n$. In particular, if $\bar{F}(x_1, \ldots, x_n)$ is absolutely continuous with density $f(x_1, \ldots, x_n)$, then the multivariate failure rate function $r(x_1, \ldots, x_n)$ is stationary in $x_1, \ldots, x_n$, and also, for $i=1,2, \ldots, n$,

$$\frac{\frac{\partial}{\partial x_i}\bar{F}(x_1, \ldots, x_n)}{\bar{F}(x_1, \ldots, x_n)}$$

is stationary in $x_1, \ldots, x_n$.

It is natural to ask if any of these conditions characterize the *LMP*. We give an affirmative answer in the following theorem.

THEOREM 5.1: Under the assumption that the first partial derivatives of $\bar{F}$ exist, $\bar{F}$ has the *LMP* if and only if

$$\frac{\frac{\partial}{\partial x_i}\bar{F}(x_1, \ldots, x_n)}{\bar{F}(x_1, \ldots, x_n)}$$

is stationary in $x_1, \ldots, x_n$ for each $i=1,2, \ldots, n$.

PROOF: If $\bar{F}$ has the *LMP*, then one direction of the proof is obvious from the remarks preceding the theorem. Now assuming the stationary conditions, for all $i=1,2, \ldots, n$,

$$\frac{\frac{\partial}{\partial x_i}\bar{F}(x_1+\Delta, \ldots, x_n+\Delta)}{\bar{F}(x_1+\Delta, \ldots, x_n+\Delta)}=\frac{\frac{\partial}{\partial x_i}\bar{F}(x_1, \ldots, x_n)}{\bar{F}(x_1, \ldots, x_n)}$$

The left-hand side of the above equation is equal to

$$\frac{\partial}{\partial x_i}\log\frac{\bar{F}(x_1+\Delta, \ldots, x_n+\Delta)}{\bar{F}(\Delta, \ldots, \Delta)}$$

and the right-hand side is

$$\frac{\partial}{\partial x_i}\log\bar{F}(x_1, \ldots, x_n)$$

so that

$$\frac{\partial}{\partial x_i}\log\frac{\bar{F}(x_1+\Delta, \ldots, x_n+\Delta)}{\bar{F}(x_1, \ldots, x_n)\bar{F}(\Delta, \ldots, \Delta)}=0.$$

Since this is true for all $i=1, 2, \ldots, n$, it follows that

$$\log \frac{\overline{F}(x_1+\Delta, \ldots, x_n+\Delta)}{\overline{F}(x_1, \ldots, x_n)\overline{F}(\Delta, \ldots, \Delta)} \tag{5.1}$$

is a function only of Δ and constant in $x_1, \ldots, x_n$. This gives that (5.1) is equal to

$$\log \frac{\overline{F}(\Delta, \ldots, \Delta)}{\overline{F}(0, \ldots, 0)\overline{F}(\Delta, \ldots, \Delta)}=0.$$

This gives the *LMP*, and the result is proven.

To see that this characterization is related to the previous material, notice that for $i=1, 2, \ldots, n$,

$$\frac{\frac{\partial}{\partial x_i}\overline{F}(x_1, \ldots, x_n)}{\overline{F}(x_1, \ldots, x_n)}$$

is stationary in $x_1, \ldots, x_n$ if and only if

$$\frac{\frac{\partial}{\partial x_i}\overline{F}(x_1+\Delta, \ldots, x_n+\Delta)}{\overline{F}(x_1+\Delta, \ldots, x_n+\Delta)}$$

is constant in Δ for each $x_1, \ldots, x_n$. This last condition is then seen to be equivalent to $\overline{F}$ being both IFR_a and DFR_a. This observation also leads to an alternate proof of the preceding theorem using our Theorem 4.1 and the results of Brindley and Thompson [5].

In the biviarate case, a related result is of some interest. Its proof is similar to that of some of the preceding results and is omitted.

THEOREM 5.2: Subject to the existence of the appropriate partial derivatives, any two of

$$\frac{\overline{F}_{x_1}(x_1, x_2)}{\overline{F}(x_1, x_2)}, \frac{\overline{F}_{x_2}(x_1, x_2)}{\overline{F}(x_1, x_2)}, \frac{\overline{F}_{x_1x_2}(x_1, x_2)}{\overline{F}(x_1, x_2)}$$

are stationary if and only if $\overline{F}$ has the *LMP*.

APPENDIX

Harris [6] was one of the first to introduce a multivariate increasing failure rate concept. He assumed (*) of Section 2 and, in addition, a type of dependence called right corner set increasing. This positive dependence concept is one of many discussed in Chapter 5, Section 4 of Barlow and Proschan [1]. Assuming any of these in conjunction with (*) would give a different monotone failure rate concept. Since this concept depends on the specific type of dependence assumed, it appears to be less important than the concepts of Brindley and Thompson discussed in the main body of the present paper.

For completeness, we discuss here the failure rate concept of Harris in the context of the present paper. Harris called his concept an increasing hazard rate.

DEFINITION (Harris): A multivariate d.f. $\overline{F}(x_1, \ldots, x_n)$ for the random variables $X_1, \ldots, X_n$ is said to have *increasing hazard rate* (*IHR*) if it is *IFR* in the sense of the definition of Section 2 and, in addition, $\overline{F}(x_1, \ldots, x_n)$ is *right corner set increasing* (*RCSI*); i.e.,

$$P\{X_1>x_1',\ldots,X_n>x_n'\mid X_1>x_1,\ldots,X_n>x_n\}$$

is increasing in $\{x_i\mid x_i\geq x_i'\}$ for each choice of $x_1',\ldots,x_n'$.

Although the original definition of *RCSI* given by Harris [6] is different than that given above, it is equivalent to it by Theorem 3.2 of Brindley and Thompson [5] (modulo certain typographical omissions). The above form is preferred since it is then possible to define a decreasing version of the concept by replacing the word increasing by decreasing. This was done by Brindley and Thompson, and $\overline{F}$ is said to have *DHR* if it is *DFR* and *right corner set decreasing* (*RCSD*), i.e. the quantity above is decreasing in $\{x_i|x_i\geq x_i'\}$ for each choice of $x_1',\ldots,x_n'$.

It is easy to see that a bivariate exponential distribution of Gumbel (see Ref. [5], p. 823) is not *IHR* but does satisfy Condition 3.1 of Section 3, so that this concept also is not related to the monotonicity concepts discussed in Section 3.

We adopt the following notation for the vectors $x=(x_1,\ldots,x_n)$ and $y=(y_1,\ldots,y_n)$:

(1) $x\geq y$ means $x_i\geq y_i$ for $i=1,\ldots,n$.

(2) $f(x)$ is increasing in x means f is increasing in each component of x.

(3) For $K\subset\{1,2,\ldots,n\}$, $\overline{K}=\{1,2,\ldots,n\}-K$ and $x_K=\{x_i|i\in K\}$, where subscripts are placed in ascending order.

As in Section 4, the following alternate hazard rate concept can be introduced.

DEFINITION: The d.f. $\overline{F}(x_1,\ldots,x_n)$ for the random variables $X_1,\ldots,X_n$ is IHR_a (DHR_a) if it is IFR_a (DFR_a) and, in addition, it is $RCSI_a$ ($RCSD_a$); i.e. for all $K\subset\{1,2,\ldots,n\}$ and each $i\in K$,

$$\frac{\overline{F}_{x_i}(x_K,x_{\overline{K}})}{\overline{F}(x_K,x_{\overline{K}})}=\frac{\frac{\partial}{\partial x_i}\overline{F}(x_K,x_{\overline{K}})}{\overline{F}(x_K,x_{\overline{K}})}$$

is increasing (decreasing) in $x_{\overline{K}}$ for all x_K.

THEOREM: A.1: Let $F(x_1,\ldots,x_n)$ be IHR_a (DHR_a). Then for each $i\in K$,

$$\frac{\overline{F}_{x_i}(x_K+\Delta,x_{\overline{K}})}{\overline{F}(x_K+\Delta,x_{\overline{K}})}$$

is decreasing (increasing) in Δ for all $x_{\overline{K}}$.

PROOF: For each $i\in K$ and for each $\Delta>0$,

$$\frac{\overline{F}_{x_i}(x_K+\Delta,x_{\overline{K}})}{\overline{F}(x_K+\Delta,x_{\overline{K}})}\leq(\geq)\frac{\overline{F}_{x_i}(x_K+\Delta,x_{\overline{K}}+\Delta)}{\overline{F}(x_K+\Delta,x_{\overline{K}}+\Delta)}\leq(\geq)\frac{\overline{F}_{x_i}(x_K,x_{\overline{K}})}{\overline{F}(x_K,x_{\overline{K}})},$$

where the first inequality follows by $RCSI_a$ ($RCSD_a$) and the second by IFR_a (DFR_a). This gives the result.

Properties (1)–(4) of Section 4 with IHR_a (DHR_a) replacing joint IFR_a (joint DFR_a) also hold. In this case (3) requires proof, but this follows by taking K in the above theorem to be the set of the indices of the requisite subset and taking $x_{\overline{K}}=0$. The proof of (4) is similar to the proof of (4) in Section 4.

The following result is a corollary to Theorem 4.1.

COROLLARY: Let $\overline{F}(x_1, \ldots, x_n)$ be *IHR* (*DHR*). Then

$$\frac{\overline{F}(x_K+\Delta, x_{\overline{K}})}{\overline{F}(x_K, x_{\overline{K}})}$$

is decreasing (increasing) in x_K for all $x_{\overline{K}}$ where $K \subset \{1, 2, \ldots, n\}$.

PROOF: We prove this only in the case that the first partial derivatives of $\overline{F}$ are assumed to exist. The proof in the general case is similar to an argument in Harris [6]. First, $\overline{F}$ is IHR_a (DHR_a) by the previous theorem. Then, by Theorem 4.1, we have that

$$\frac{\overline{F}_{x_i}(x_K+\Delta, x_{\overline{K}})}{\overline{F}(x_K+\Delta, x_{\overline{K}})}$$

is decreasing (increasing) in Δ for each $i \in K$ and for all $x_{\overline{K}}$. Then, as in the proof of Theorem 4.1,

$$\frac{\partial}{\partial x_i} \frac{\overline{F}(x_K+\Delta, x_{\overline{K}})}{\overline{F}(x_K, x_{\overline{K}})} \leq (\geq) 0$$

and so the result holds.

THEOREM A.2: Subject to the existence of the first partial derivatives of the d.f. $\overline{F}$. $\overline{F}$ is $RCSI_a$ ($RCSD_a$) if and only if F is *RCSI* (*RCSD*).

PROOF: Let $K \subset \{1, 2, \ldots, n\}$. Then $\overline{F}$ is $RCSI_a$ ($RCSD_a$) means that

$$\frac{\overline{F}_{x_i}(x_K, x_{\overline{K}})}{\overline{F}(x_K, x_{\overline{K}})}$$

is increasing (decreasing) in $x_{\overline{K}}$ for all $i \in K$. Then for all $i \in K$ and $x'_{\overline{K}} > x_{\overline{K}}$, we have

$$\frac{\overline{F}_{x_i}(x_K, x_{\overline{K}})}{\overline{F}(x_K, x_{\overline{K}})} \leq (\geq) \frac{\overline{F}_{x_i}(x_K, x'_{\overline{K}})}{\overline{F}_{x_i}(x_K, x'_{\overline{K}})}.$$

This implies that for all $i \in K$ and for $x'_{\overline{K}} > x_{\overline{K}}$,

$$0 \leq (\geq) \frac{\partial}{\partial x_i} \frac{\overline{F}(x_K, x'_{\overline{K}})}{\overline{F}(x_K, x_{\overline{K}})},$$

and so for $x'_{\overline{K}} > x_{\overline{K}}$,

$$\frac{\overline{F}(x_K, x'_{\overline{K}})}{\overline{F}(x_K, x_{\overline{K}})}$$

is increasing (decreasing) in x_K. This gives that

$$P(X > x' | X > x)$$

is increasing in $\{x_i | x_i \geq x'_i\}$ and is *RCSI* (*RCSD*). The argument is reversible and so the result is proven.

COROLLARY: Under the assumptions of Theorem A.2 on $\overline{F}$, $\overline{F}$ is IHR_a (DHR_a), if and only if $\overline{F}$ is *IHR* (*DHR*).

PROOF: Apply Theorems 4.1 and A.2.

REFERENCES

[1] Barlow, R. E., and F. Proschan, *Statistical Theory of Life Testing: Probability Models.* (Holt, Rinehart and Winston, Inc., New York, 1975).

[2] Basu, A. P., "Bivariate Failure Rate," Journal of the American Statistical Association, *66*, 103–104 (1971).

[3] Block, H. W., "Continuous Multivariate Exponential Extensions," in *Reliability and Fault Tree Analysis*, R. E. Barlow, J. B. Fussell, and N. D. Singpurwalla, (Editors), pp. 285–306, (SIAM, Philadelphia, Pennsylvania, 1975).

[4] Block, H. W., and A. P. Basu, "A Continuous Bivariate Exponential Extension," Journal of the American Statistical Association, *69*, 1031–1037 (1974).

[5] Brindley, E. C., Jr., and W. A. Thompson, Jr., "Dependence and Aging Aspects of Multivariate Survival," Journal of the American Statistical Association, *67*, 822–830 (1972).

[6] Harris, R., "A Multivariat Defineition for Increasing Hazard Rate Distribution Functions," Annals of Mathematical Statistics, *41*, 713–717 (1970).

[7] Johnson, N. L., and S. Kotz, "A Vector Multivariate Hazard Rate," Journal of Multivariate Analysis, *5*, 53–66 (1975).

[8] Marshall, A. W., "Some Comments on the Hazard Gradient," Stochastic Processes and their Applications *3*, 293–300 (1975).

[9] Marshall, A. W., and I. Olkin, "A Multivariate Exponential Distribution," Journal of the American Statistical Association, *62*, 30–44 (1967).

[10] Puri, P. S., "On a Property of Exponential and Geometric Distributions and Its Relevance to Failure Rate," Sankhya, Series A, *35*, 61–68 (1973).

[11] Puri, P. S., and H. Rubin, "On a Characterization of the Family of Distributions with Constant Multivariate Failure Rates," Annals of Probability, *2*, 738–740 (1974).

AN ANALYTICAL MINEFIELD EVALUATION MODEL WITHOUT SPACE AVERAGES

K. M. Mjelde

SHAPE Technical Centre
The Hague, Netherlands

ABSTRACT

The paper describes an approach to the evaluation of the effectiveness of a minefield in terms of the number of mines that are detonated by a convoy of sweepers and ships and the corresponding number of vessels that are immobilized. The positions of the mines and the tracks of the vessels are assumed to be known, which means that the evaluation measures are dependent on a large number of disjoint events, each event being the immobilization of particular vessels by particular mines. This may render combinatorial methods computationally infeasible, but by introducing approximations in the assumptions, the difficulty can be overcome, specifically by modelling the arrival of each individual vessel in the neighborhood of a mine by an inhomogeneous Poisson stream for which the arrival rate is non-zero only over a short time interval. The plausibility of the approach is supported by results of a critical-event simulation model.

INTRODUCTION

The discussion given in this paper concerns the characteristics of minefields laid in the path of transiting ships. Minesweepers pass to and fro across the fields several times in an attempt to clear a channel prior to the arrival of ships. The term "vessel" is used to refer to a ship or a sweeper. The mines are fitted with a count device such that a fixed number of actuations caused by passing vessels has to take place before a mine detonates. The situation is analysed in terms of the number of mines that detonate and the losses inflicted on the vessels.

Several authors have considered this problem; the papers of Conolly [1] and Hill [2] may be mentioned, but may not be readily available to most readers. The present paper extends previous work by the explicit introduction of:

1. **the geographical positions of the mines,**
2. **the tracks and relative positions of the vessels in the convoy, and**
3. **the following considerations that were found necessary for practical applications:**
   - **(1) a mine usually has an intercount dormant period (the minimum time interval between successive actuations of the mine);**

(2) a single sweeper may actuate a mine several times;
(3) vessels may make multiple crossings of the field;
(4) vessels are subjected to navigation errors;
(5) the reliability of a mine could be uncertain;
(6) mines frequently have arming delays (initial period of dormancy).

Factors 1 and 2 imply that the probability of a mine being actuated by a passing vessel and the conditional probability of immobilization given detonation can be specified in terms of the relative positions of vessel and mine. These probabilities of actuation and immobilization may be any function of the distance between the mine and the track of the vessel. In Refs. [1] and [2], an average actuation probability is used as input to the basic mathematical equations describing the losses of vessels and mines. The spatial averaging with respect to the positions of the mines and the tracks of the vessels is performed before the losses are calculated. The result is not identical to that obtained by first calculating the losses for given mine positions and vessel tracks and then taking the spatial average value, which is a more correct procedure. Furthermore, the introduction of an average actuation probability for each mine type may be of little use if there are only a few mines of each type near the channel cleared by the minesweepers. Another difficulty is revealed by considering a convoy with a very large number of vessels crossing an extensive minefield containing mines that have zero and nonzero actuation probabilities with respect to the vessels in the convoy. The use of an average actuation probability for each type of mine and vessel, on the assumption that the positions of the mines are uniformly distributed in the field, implies that all mines of each type have the same nonzero probability of being actuated by a vessel. A very large convoy could thus theoretically cause all mines to detonate, whereas in reality only the mines in a narrow channel would be detonated. An obvious way to circumvent this difficulty is to discard all the mines outside the actuation ranges of the vessels, but because actuation ranges vary from mine to mine and from vessel to vessel, there is difficulty in deciding which mines should be discarded and which should not. A mine that has an actuation range much greater than normal may prove to be deadly for one of the important or valuable ships in the convoy; on the other hand, choosing too wide a channel results in too many losses. The introduction of the inputs 1 and 2 resolves these difficulties.

Extensive descriptions of the model and the corresponding computer program are given in Ref. [4]. Numerical comparisons based on a detailed and much more aggregated critical-event simulation model support the theory. Consequently, the analytical model's advantage of being much less time-consuming recommends its use in place of simulation models.

In the present paper, the description of the situation to be analysed and of the associated mathematical parameters is followed by a model for the losses of sweepers, ships, and mines. The verification of the model is then described, and several extensions are introduced. Finally, a plausible numerical example with fictitious data is given for illustration purposes.

THE SITUATION

For model design, it is assumed that merchant ships and combat ships cross an ocean region at the same speed and in the same direction. The paths of the ships cross a minefield, where the ships become vulnerable to damage as a consequence of mine detonations. Minesweepers pass to and fro across the minefields several times in an attempt to detonate as many mines as possible, thereby clearing a channel through the fields prior to the arrival of the ships. The arriving ships then cross the field along this channel, possibly preceded by minesweepers.

A vessel (ship or sweeper) has prescribed probabilities of actuating mines that are located sufficiently close to its track. When a mine is actuated, its "count" is raised by one. When the count has increased by the number of times fixed by the count setting, the mine detonates and has a prescribed probability of destroying the actuating vessel, depending on its distance from the vessel at the instant of detonation. Various levels of damage can be defined, but the one used in this paper is the level which causes immobilization of the vessel.

The specific problem addressed is quantification of the total numbers of mines detonated and vessels destroyed.

PARAMETERS

Vessels and mines are identified by subscripts i and j, respectively. The following parameters are used:

I : the total number of vessels,

J : the total number of mines,

m_j : the count setting of mine j, $j=1, \ldots, J$,

α_{ji} : the conditional probability of vessel i actuating mine j, given that the vessel approaches the mine and that the mine is active, $j=1, \ldots, J$ and $i=1, \ldots, I$,

P_{ji} : the probability of vessel i being immobilized, given that it detonates mine j, $j=1, \ldots, J$ and $i=1, \ldots, I$.

Vessels and mines are numbered from 1 to I and from 1 to J, respectively, in the sequence of their projections on a line parallel to the direction of motion of the vessels, and such that vessel number 1 bypasses mine number 1 first. When a mine is actuated, its count is raised by one, and after m_j actuations the mine detonates. No restrictions are imposed on the submodels that are used to obtain the actuation and damage probabilities α_{ji} and P_{ji}; these submodels will depend on the tracks of the vessels and the positions of the mines.

In the mathematical derivations, it is assumed that any chosen vessel in the convoy can actuate a mine only once; it is then demonstrated how the equations can be modified to allow for multiple actuations (for example, by a sweeper).

THE MODEL

A vessel which is immobilized by a mine can obviously not actuate any subsequent mine in the field. Consequently, for each mine j and vessel i, the probability that vessel i detonates mine j depends on whether any of the vessels $\{k{:}1\leq k\leq i\}$ has been immobilized by one of the mines $\{m, 1<m<j-1\}$. The implication is that the probability that a given number of vessels are immobilized depends on a large number of disjoint events. In fact, a combinatorial model that adds up the probabilities of these events may become computationally impractical unless approximate evaluation methods can be introduced.

A simplification results if it is noted that the probability of vessel i being immobilized by mine j, given which vessels out of $1, \ldots, (i–1)$ were destroyed by the mines $1, \ldots, (j–1)$, can be expressed in terms of the binomial probability distribution, if the actuation and destruction probabilities for all mines vs all vessels are identical. This means that these probabilities are described by their spatial average values. However, for reasons previously given, this approach is not used in this paper. Furthermore, when the number of vessels immobilized cannot be considered small, the basic combinatorial problem described above in determining the probability distribution of the total number of vessels immobilized is unsolved.

The method used in this paper is based on the introduction of approximations in the model assumptions: vessels are assumed to arrive at the minefield in an inhomogeneous Poisson stream with rate $\mu(\theta)$ for $\theta \geq 0$. Vessel i is associated with a time interval $[a_i, b_i]$ such that

$$0<a_i<b_i<a_{i+1}<b_{i+1},$$

showing that vessel i arrives prior to vessel $(i+1)$. Furthermore,

$$\int_{a_i}^{b_i} \mu(\theta)\,d\theta=1$$

and

$$\mu(\theta)=0 \text{ for } \theta \notin \bigcup_{i=1}^{I} [a_i, b_i].$$

Then vessel i is described by the rate $\mu(\theta)$ over the time interval $[a_i, b_i]$. It turns out that the results of this paper do not depend on the choice of the values of a_i and b_i if the above requirements are satisfield.

At this point, it is important to describe how the vessels actuate the mines and how mine detonation influences the progress of the vessels through the field. To this end, let $\mu_{ji}(\theta)$ be the rate in an inhomogeneous Poisson stream describing the arrival of vessel i at mine j. Then

$$\mu_{1i}(\theta)=\begin{cases} \mu(\theta), & \text{for } \theta \epsilon [a_i, b_i] \\ 0, & \text{otherwise.} \end{cases}$$

The $\mu_{ji}(\theta)$ are calculated recursively by (4) that is derived below. It follows that $\mu_{ji}(\theta)>0$ only for $\theta \epsilon [a_i, b_i]$. Mine j is actuated in the time interval $[\theta, \theta+\Delta\theta]$ with probability $\mu_{ji}(\theta)\alpha_{ji}\Delta\theta$, if $\theta \epsilon [a_i, b_i]$. Actuations in different time intervals $\Delta\theta$ are independent. As a check on the assumptions, note that the probability that vessel i actuates mine 1 is given by

$$\int \mu_{1i}(\theta)\alpha_{1i}\,d\theta=\alpha_{1i},$$

as it should.

If $p_{jk}(\theta)$ is the probability that mine j receives k actuations before time θ, then a set of differential-difference equations for these probabilities can be written, with the solution

$$p_{jk}(\theta)=\frac{1}{k!} E_j(\theta)^k \exp(-E_j(\theta)) \tag{1}$$

where

$$E_j(\theta)=\sum_{i=1}^{I} \alpha_{ji} \int_0^{\theta} \mu_{ji}(v)\,dv. \tag{2}$$

Putting $k=m_j-1$ gives the probability $p_j(\theta)$ that the mine detonates if it is actuated once more at time θ:

$$p_j(\theta)=p_{j(m_j-1)}(\theta). \tag{3}$$

Note that when all α_{ji} and $\mu_{ji}(v)$ for $i=1, \ldots, I$ are equal, (1) reduced to the well-known Poisson approximation of the binomial distribution.

Remember that $\mu_{1i}(\theta) i=1, \ldots, I$ are given. Assuming that the $\mu_{ji}(\theta)$, $i=1, \ldots, I$ are known for a particular value of j, the $\mu_{j+1,i}(\theta)$, $i=1, \ldots, I$ are obtained as follows:

$$\mu_{j+1,\,i}(\theta)=\mu_{ji}(\theta)\,(1-K_{ji}(\theta)) \tag{4}$$

where

$$K_{ji}(\theta)=P_{ji}\alpha_{ji}p_j(\theta). \tag{5}$$

In explanation of (4) and (5), note that $\mu_{ji}(\theta)\Delta\theta\alpha_{ji}p_j(\theta)$ is the probability that vessel i detonates mine j in $\Delta\theta$. Then the vessel is destroyed with probability P_{ji}; the arrival rate is modified accordingly.

The total losses can conveniently be expressed in terms of the probabilities q_{ji} that each ship i detonates mine j. Then

$$q_{ji}=\int_{a_i}^{b_i}\alpha_{ji}\mu_{ji}(\theta)p_j(\theta)\,d\theta. \tag{6}$$

If we write

$$C_{ji}=\int_{a_i}^{b_i}\mu_{ji}(\theta)\,d\theta,$$

it follows that for $\theta\epsilon[a_i, b_i]$,

$$E_j(\theta)=\sum_{n=1}^{i-1}\alpha_{jn}C_{jn}+\alpha_{ji}\int_{a_i}^{\theta}\mu_{ji}(v)\,dv,$$

where the sum on the right-hand side is zero for $i=1$. Using (1), (2), (3), and (6), we obtain the following expressions for the detonation probabilities:

$$q_{j1}=1-\sum_{k=0}^{m_j-1}\frac{1}{k!}\,(E_{j1})^k\exp\,(-E_{j1}) \tag{7}$$

$$q_{ji}=\sum_{k=0}^{m_j-1}\frac{1}{k!}\,[(E_{j,\,i-1})^k\exp\,(-E_{j,\,i-1})-(E_{ji})^k\exp\,(-E_{ji})];\ 2\leq i\leq I, \tag{8}$$

where

$$E_{ji}=E_{j,\,i-1}+\alpha_{ji}C_{ji} \tag{9}$$

and

$$E_{j0}=0. \tag{10}$$

Integrating (4) and using (5) and (6) gives

$$C_{j+1,\,i}=C_{ji}-P_{ji}q_{ji}. \tag{11}$$

The values of q_{ji} are calculated recursively; when q_{ji} is known for $i=1, \ldots, I$, the $C_{j+1,\,i}$ are given by (11) and the $q_{j+1,\,i}$ by (7) and (8), the recursion starting with $C_{1i}=1$ and the calculable q_{1i}.

The probabilities x_{ji} that each vessel i is destroyed by mine j are given by

$$x_{ji}=P_{ji}q_{ji}\ \ j=1, \ldots, J \text{ and } i=1, \ldots, I, \tag{12}$$

and the probability-generating function $F_j(z)$ of the number of vessels destroyed by mine j is

$$F_j(z)=\left(\sum_{i=1}^{I}x_{ji}\right)z+\left(1-\sum_{i=1}^{I}x_{ji}\right),$$

since

$$\sum_{i=1}^{I} x_{ji}$$

is the probability that some vessel is destroyed by mine j.

As a check, note that if the quantity $C_{j+1,\,i}$ in (11) is interpreted as the probability that vessel i arrives at mine $(i+1)$, then (11) expresses this probability as the corresponding probability C_{ji} minus the probability x_{ji} given in (12), in accordance with intuition.

Since the vessels are described by Poisson streams, the probability generating function of the total number of vessels that are destroyed is obtained by

$$F(z)=\prod_{j=1}^{J} F_j(z), \tag{13}$$

from which the expected value x and variance σ^2 can be derived:

$$x=\sum_{j=1}^{J}\sum_{i=1}^{I} x_{ji} \tag{14}$$

$$\sigma^2=\sum_{j=1}^{J}\left\{\sum_{i=1}^{I} x_{ji}(1-x_{ji})-2\sum_{i_1<i_2} x_{ji_1}x_{ji_2}\right\} \tag{15}$$

All vessels are given the value "1" in the previous equations. However, the loss of a ship with troops may be more serious than the loss of a sweeper. This is described by the introduction of the combat values a_i, $i=1, \ldots, I$ of the vessels. The combat values can be introduced in (14) and (15) by replacing x_{ji} in (14) by $a_i x_{ji}$, and the terms $x_{ji}(1-x_{ji})$ and $x_{ji_1}\ x_{ji_2}$ in (15) by $a_i^2\ x_{ji}(1-x_{ji})$ and $a_{i_1}\ a_{i_2}\ x_{i_1}\ x_{i_2}$, respectively. Furthermore, the losses within a particular group of ships, indicated by the subscripts i as belonging to a subset G of $\{1, \ldots, I\}$, can be obtained by restricting the summations in (14) and (15) to those i for which $i \epsilon G$.

The number of mines that are detonated can be obtained by putting $P_{ji}=1$ for all $1\leq j\leq J$ and $1\leq i\leq I$ in (12) and using the (13)–(15).

EXPERIMENTAL VERIFICATION

The plausibility of the assumptions on which the theoretical model has been founded is supported by the results of critical event simulations. The probability distribution $\{f_n\}$ of the total losses of vessels was found from the probability generating function $F(z)$ given in (13) by using a Fast Fourier inversion procedure. In order to use the computer program of Ref. [5], one determines an integer m such that $2^{m-1}<J\leq 2^m$ and writes $M=2^m$. Next, (13) is used to obtain the Fourier transform:

$$C_r=\frac{1}{M}F\left(\exp\left(-i\frac{2\pi r}{M}\right)\right), \text{ where } r=0, 1, 2, \ldots, M-1,$$

from which the probability distribution $\{f_n\}$ is obtained by the inversion formula

$$f_n=\sum_{r=0}^{M-1} C_r \exp\left(i\frac{2\pi rn}{M}\right), \text{ where } n=0, 1, 2, \ldots, J.$$

Note that $f_n=0$ for $n>J$.

Regarding the probability distribution and the expected value of the losses of vessels, it was found that the results of the simulation model oscillated around the corresponding results from the analytical model. These experimental errors did decrease and the simulation results clustered more tightly around the analytical model results when the number of replications of the simulation was increased. With 100 replications, the sampling errors were greater than any systematic error in the probability distribution and the expected losses of vessels as calculated by the analytical model; any systematic errors in the expected values were almost negligible for all the parameter values considered (see Ref. [3]).

The standard deviations of the losses of vessels as calculated by the two models also agreed closely, except when the number of vessels was smaller than about 30 (in this case, the standard deviation obtained from the analytical model could be up to 25% higher than that calculated by the simulation model).

EXTENSIONS

The practical value of the model can be extended by taking explicit account of the following factors:

(a) the intercount dormant periods of the mines,

(b) the possibility of a sweeper actuating a mine several times,

(c) multiple crossings of the field by sweepers and ships,

(d) navigation errors of the vessels,

(e) the reliability of the mines, and

(f) arming delays of the mines.

A brief discussion of the implications of these factors is now given.

Intercount Dormant Periods

These make the field less vulnerable to sweeping and are described by the introduction of the following parameters:

δ_j: The intercount dormant period of mine j; this is the minimum time interval between successive actuations of the mine.

β_i: The distance between vessel i and the next following vessel, measured by the time between the instants when these vessels cross a line perpendicular to their direction of motion, $i=1, \ldots, I$.

DEFINITION:

d_{ji}: The probability that vessel i actuates mine j, given that the vessel has not been immobilized by any previously encountered mine. (Not conditional upon the mine being active.)

For each mine j, the probabilities d_{ji} for $i=1, \ldots, I$ are calculated recursively, starting with $i=1$ and increasing i in steps of 1:

$$d_{ji}=\Big(1-\sum_{k \epsilon K_{ji}} d_{jk}\Big)\alpha_{ji},\ i=1, \ldots, I, \tag{16}$$

where K_{ji} is the set of all vessel subscripts $k<i$ such that the time distance between the vessels k and i is smaller than the intercount dormant period of mine j; that is,

$$K_{ji}=\left\{k|k<i \text{ and } \sum_{n=k}^{i-1} \beta_n<\delta_j\right\}.$$

In (16), it is assumed that

$$\sum_{k\epsilon K_{ji}} d_{ji}=0 \text{ when } K_{ji}=\Phi.$$

In explanation, note that the events that the various vessels $k \epsilon K_{ji}$ actuate mine j are mutually disjoint and that each of these events prevents actuation of mine j by vessel i. The term in brackets in (16) gives the probability that none of the vessels $k\epsilon K_{ji}$ actuates mine j, which implies that the mine is active upon arrival of vessel i and is consequently actuated with probability α_{ji}.

The probabilities d_{ji} are now used in the equations of the analytical model, instead of the α_{ji}. Basically, the intercount dormant periods are introduced by their effects on the arrival rates of vessels, if one uses (4) and (5).

Multiple Actuations

Multiple actuations of mine j by sweeper i can be described by a parameter A_{ji} giving the expected number of actuations, under the condition that at least one actuation did occur. Thus, each term $\alpha_{jn}C_{jn}$ (or $d_{jn}C_{jn}$) of (7)–(10) for q_{ji} has to be replaced by $\alpha_{jn}C_{jn}A_{jn}$.

Multiple Crossings

The equations can be extended to the case where a convoy or a collection of sweepers passes several times to and fro across the minefield; for example, when the escort of a convoy which has just come from port through a narrow channel is required to escort a newly arrived group of ships back through the channel to port.

The quantities E_{ji} given by (9) and (10) and used in (7) and (8) can be interpreted as actuation potentials that are accumulated on the mines by the vessels. In the first crossing, the initial value of the actuation potential is zero, $E_{j0}=0$, and after all of the vessels have crossed the minefield the actuation potential of mine j is given by E_{jI}. Similarly, the quantities C_{ji} that are calculated recursively by using equation (11) are the probabilities that each vessel i arrives at each mine j, and $C_{L+1,i}$ are the probabilities that each vessel i exits from the minefield without suffering immobilization.

If each vessel returns along the same track that it used in the first crossing, then a mine numbered j in the return crossing had the number $(L-j+1)$ in the forward crossing. If we use a left-hand superscript beside a quantity to denote the number of the crossing, the second crossing is described by putting

$$^2E_{ji}={}^2E_{j,i-1}+{}^2\alpha_{ji}{}^2C_{ji}$$

with

$$^2E_{j0}={}^1E_{L-j+1,I}=E_{L-j+1,I}.$$

The $^2q_{ji}$ are calculated by (7) and (8), with $^2E_{ji}$ replacing E_{ji} and $^2C_{ji}$ replacing C_{ji}, where

$$^2C_{j+1,i}={}^2C_{ji}-{}^2P_{ji}{}^2q_{ji}$$

and

$$^2C_{1i}={}^1C_{L+1,i}=C_{L+1,i}.$$

Also

$${}^2\alpha_{ji}={}^1\alpha_{L-j+1,\,i}$$

and

$${}^2P_{ji}={}^1P_{L-j+1,\,i}.$$

However, these equations must be modified if the tracks used in the two crossings are related by some more general transformation such as a parallel translation or rotation. In such cases, the mines must be reordered according to their projections on the new tracks, and the probabilities ${}^2\alpha_{ji}$ and ${}^2P_{ji}$ must be recalculated in terms of the new distances between vessels and mines.

The total losses of vessels in the second crossing are calculated by using (12)–(15) and letting q_{ji} in (12) be replaced by ${}^2q_{ji}$. Following crossings are treated in an analogous way.

Navigation Errors

Navigation errors are introduced in the calculation of the probabilities α_{ji} that ship i actuates mine j, given that the ship approaches the mine and that the mine is active. Let D_{ji} be the distance between the planned track of the vessel i and mine j; let $a_{ji}(x)$ be the probability that the vessel actuates the mine, given the distance x to its track; let $n_{ji}(x)$ be the probability distribution of navigation errors measured by the distancc of the vessel i from its planned track when it passes mine j. Then

$$\alpha_{ji}=\int_{-\infty}^{\infty} a_{ji}(D_{ji}-x)\,n_{ji}(x)\,dx.$$

Note that the above approach slightly violates the principle that the losses should be determined for given tracks of the vessels before taking spatial averages. However, the mean tracks of the vessels are retained in the above description such that the dependences between the tracks of the vessels are essentially retained. Averaging with respect to the mean tracks of vessel can then be performed afterwards.

Reliability of Mines

The reliability of a mine can be described by the probability that it functions properly; these probabilities are used to multiply the corresponding actuation and damage probabilities α_{ji} and P_{ji}.

Arming Delays

If a mine j is armed during the transition of a convoy through the field, this can be described by replacing the actuation probabilities α_{ji} by the probabilities α^*_{ji}, where $\alpha^*_{ji}=0$ until the mine is armed; if arming occurs when vessel i_0 passes the mine, then $\alpha^*_{ji}=\alpha_{ji}$ for $i\geq i_0$.

NUMERICAL EXAMPLE

A simple numerical example is now given. The convoy consists of 30 ships following each other in a single row, crossing a field of 50 mines arranged in 10 lines of 5 mines each, the lines being perpendicular to the direction of motion of the ships. The ships and mines are numbered according to the previously described rules. If we consider the first ship and the mines with numbers 1, 2, 3, 4, and 5 in the first line of the field, the actuation and immobilization probabilities are as follows:

$$d_{11}=0.10,\ d_{21}=0.20,\ d_{31}=0.40,\ d_{41}=0.20,\ d_{51}=0.10;$$
$$P_{11}=0.05,\ P_{21}=0.10,\ P_{31}=0.20,\ P_{41}=0.10,\ P_{51}=0.05.$$

Each actuation and immobilization probability increases by 0.01 when the sequence number of the ship increases by 1, such that

$$d_{1i}=d_{5i}=0.09+0.01.i;\ d_{2i}=d_{4i}=0.19+0.01.i;\ d_{3i}=0.39+0.01.i;$$
$$P_{1i}=P_{5i}=0.04+0.01.i;\ P_{2i}=P_{4i}=0.09+0.01.i;\ P_{3i}=0.19+0.01.i;\ (i=1,\ .\ .\ .,\ 30).$$

The actuation and immobilization probabilities within each of the 10 lines of mines are identical. If we write $j=5(n\text{–}1)+m$ for integers n and m where $1\leq n\leq 10$ and $1\leq m\leq 5$, it follows that

$$d_{ji}=d_{mi} \text{ and } P_{ji}=P_{mi} \text{ for } j=1,\ .\ .\ .,\ J,\ i=1,\ .\ .\ .,\ I.$$

The count settings of the mines in a line are all equal; it is 10 in the first line, 9 in the second line and so on, decreasing by 1 from one line to the next, so that

$$m_j=10-(n-1).$$

The numerical values of the probabilities given above reflect the fact that the track of the vessels passes directly above mines with numbers $j=5(n\text{–}1)+3$ for $n=1,\ .\ .\ .,\ 10$; the other mines are located some distance away from the track. For each ship and mine, the actuation and immobilization probabilities decrease when the distance between the ship and the mine increases.

The probability distribution of the number of ships that are immobilized and the corresponding expected value and standard deviation are given in the Tables 1 and 2, respectively. The losses are calculated by the analytical model and, as a check, by a critical event simulation model. Tables 1 and 2 designate as 1, 2, and 3, respectively, those numerical results obtained by the analytical model, the critical event simulation model with 100 replications, and the simulation model with 1000 replications.

TABLE 1. *The Probability Distribution*

No. of ships lost		5	6	7	8	9	10	11	12	13	14	15	16	17	18
Model	1	0. 01	0. 03	0. 05	0. 08	0. 11	0. 14	0. 14	0. 13	0. 11	0. 08	0. 05	0. 03	0. 02	0. 01
	2	0. 02	0. 03	0. 11	0. 07	0. 20	0. 12	0. 11	0. 09	0. 10	0. 07	0. 03	0. 04	0. 00	0. 01
	3	0. 02	0. 04	0. 06	0. 10	0. 14	0. 15	0. 15	0. 12	0. 10	0. 06	0. 03	0. 01	0. 01	0. 00

TABLE 2. *The Expected Value and Standard Deviation*

Parameter		Expected value	Standard deviation
Model	1	10. 12	2. 79
	2	10. 45	2. 75
	3	10. 38	2. 56

The numerical results support the statements made earlier that the analytical model provides a satisfactory substitute for simulation. It is also less demanding in computer time. While such comments, based on numerical comparison, cannot be conslusive, they do support strongly the hypothesis concerning plausibility of the approximate analytical treatment of a model whose details are manifold and complicated.

REFERENCES

[1] Conolly, B. W., "Mathematics of Military Operational Research, Part 2: A Minefield Model," SACLANTCEN Memorandum SM-21, (1973).

[2] Hill, W., "Determining the Variable Effect of Mining on Convoys," Journal of the Royal Naval Scientific Service, pp. 15-20, (1966).

[3] Mjelde, K. M., "The Applicability of a Minefield Evaluation Model with Poisson Approximations," SHAPE Technical Centre TM-469, (1975).

[4] Mjelde, K. M., and F. M. Blaauw, "Extensions of a Stochastic Model for the Evaluation and Optimization of Minefields," SHAPE Technical Centre, TM-444, (1975).

[5] Urich, M. L., "Fast Fourier Transforms Without Sorting," IEEE Transactions on Audio and Electro-acoustics, *AU-17*, 170-172, (1969).

EXPLICIT STEADY STATE SOLUTIONS FOR A PARTICULAR $M^{(X)}/M/1$ QUEUEING SYSTEM

George L. Jensen

University of Tennessee
Knoxville, Tennessee

and

Albert S. Paulson
Pasquale Sullo

Rensselaer Polytechnic Institute
Troy, New York

ABSTRACT

An explicit steady state solution is determined for the distribution of the number of customers for a queueing system in which Poisson arrivals are bulks of random size. The number of customers per bulk varies randomly between 1 and m, m arbitrary, according to a point multinomial, and customer service is exponential. Queue characteristics are given.

1. INTRODUCTION

In numerous queueing theoretic applications, arrivals at a service facility are bulks, i.e. sets of possibly more than one customer, while service is rendered, as usual, to individual customers. In these situations, moreover, it is generally unrealistic and misleading to treat the bulks as being of a specified constant size. Rather, bulk size will be a random variable. Indeed, in many applied settings, the arriving entity may be a physical unit, but represents to the server a randomly sized "bulk" of tasks, e.g., airplanes arrive at a maintenance facility each with a random number of required repairs.

The problem of randomly sized bulks has been treated in some generality in a fundamental paper by Gaver [2]. Also, in recent books by Gross and Harris [4] and Kleinrock [8], the difference equations for the steady state probability distribution of the number of customers in the system is derived for $M^{(X)}/M/1$, where $M^{(X)}$ refers to the "Markovian" arrival of bulks of random size X. Naturally, these differ nce equations depend on the probability distribution of X. In this paper, we give an explicit representation of the steady state probability distribution of the number of customers for a particular distribution of X which seems appropriate for many applications by virtue of its considerable flexibility. In so doing, we generalize the results of Harris [5], a summary of which appears in Ref. [4], section 4.1.

2. DESCRIPTION OF THE MULTIPLE POISSON BULK ARRIVAL QUEUEING SYSTEM

Assume that bulks of size j arrive at a service facility according to a Poisson process $\{N_j(t):t\geq 0\}$ with intensity $\lambda_j\geq 0, j=1, \ldots, m$, and that these m Poisson processes are mutually independent. Note that we are placing an upper limit of m on bulk size. Under these assumptions, the superposed process $\{M(t):t\geq 0\}$ with $M(t)=N_1(t)+\ldots+N_m(t)$ is a Poisson process with intensity

$$\sum_{j=1}^{m}\lambda_j=\lambda,$$

say. $M(t)$ represents the number of bulks arriving in $(0, t]$ and λ is the mean rate of bulk arrival. The process $\{N(t):t\geq 0\}$, where $N(t)=N_1(t)+2N_2(t)+\ldots+mN_m(t)$, is a compound Poisson process in which $N(t)$ is the number of individual customers arriving in $(0, t]$. Clearly,

$$E[N(t)]=\sum_{j=1}^{m} j\lambda_j t$$

and, thus,

$$\sum_{j=1}^{m} j\lambda_j$$

is the mean rate of customer arrival. Because the arriving bulks are governed by generally different Poisson processes, we shall refer to this queueing system as the *multiple Poisson bulk arrival system.*

Note that due to the so-called "preservation under random selection" property of mutually independent, superposed Poisson processes, the arrivals in $\{M(t):t\geq 0\}$ are bulks of random size X, where X has a point multinomial distribution with $P(X=j)=\lambda_j/\lambda$, $j=1, \ldots, m$ (see Ross [9] p. 123).

As previously mentioned, service is rendered to individual customers and not to bulks as such. We shall assume service to be accomplished by a single server who operates according to a Poisson process with intensity μ. Thus, the time expended per customer is an exponential random variable with mean $1/\mu$.

To fix ideas at this point, we can refer to the following simple model. Sheets of plywood coming off a processing line are rejected if they contain more than m defects, but are sent on to a repair facility if they contain between one and m defects. The sheets are envisaged to arrive at the repair facility according to a Poisson process with intensity λ, while the probability that a sheet chosen at random contains j defects is $\lambda_j/\lambda, j=1, 2, \ldots, m$. At the repair facility, a single repairman eliminates defects individually, according to a Poisson process with intensity μ.

Now let p_k be the steady state probability that the bulk arrival system contains k customers (k defects awaiting repair in the plywood example). Harris [5] has analyzed the present model and specified $\{p_k\}$ for $m=2$ only. We shall give an explicit expression for p_k, $k=0, 1, 2, \ldots$ for arbitrary (but finite) m.

The multiple Poisson bulk arrival system model, along with its explicit solution, seems particularly useful in and readily applicable to the analysis of real situations. The model assigns m

parameters to the arrival process, and consequently, it is reasonable to assume that it would provide an excellent fit to actual experience as opposed, for example, to the assumption of geometric or Poisson bulk size, each of which adds only one parameter to the arrival process. Yet, the number of parameters incorporated does not pose a difficult problem statistically, for λ can be estimated by the sample mean bulk arrival rate and λ_j/λ can be estimated by the empirical frequency of bulks of size j, $j=1, \ldots, m$. Additionally, we have the advantage of the explicit expressions for the. p_k's, which are more convenient from the aspect of statistical control than recursive generation, Examples of the application of this model for $m=2$ are given in Harris [5] (see also Ref. [4], Ex. 4.1 p. 151–154).

3. THE DIFFERENCE EQUATION FORM OF THE STEADY STATE DISTRIBUTION

The birth-death equations for the multiple Poisson bulk arrival system can be derived from the postulates associated with the various Poisson processes operative in the model. These will yield the steady state solution in the form of difference equations by the usual methods. This procedure was actually employed in Jensen [7]. Recently, however, by a method based on visual inspection of the so-called state-transition-rate diagram, Kleinrock [8] (see also Ref. [4], section 4.1) derived the steady state equations for the bulk arrival process with a general bulk size distribution. If X is the random bulk size, then for the case treated here

$$g_j \equiv P(X=j) = (\lambda_j/\lambda)\,\delta(m-j),\ j=1, 2, \ldots, \tag{1}$$

where

$$\delta(x)=1 \text{ if } x\geq 0,$$
$$=0 \text{ if } x<0.$$

Substituting g_j given by (1) into Eq. (4.44) and (4.45) of Ref. [8], p. 135, we obtain directly the steady state difference equations

$$(\lambda+\mu)p_k = \mu p_{k+1} + \sum_{i=0}^{k-1} p_i \lambda_{k-i}\,\delta(m-k+i),\ k\geq 1, \tag{2}$$

$$\lambda p_0 = \mu p_1. \tag{3}$$

Define

$$\rho_j = \sum_{k=j}^{m} \lambda_k/\mu, \qquad j=1, \ldots, m. \tag{4}$$

It can be seen after routine manipulations that (2) and (3) can be condensed to the recursive re lationship

$$p_{k+1} = \sum_{j=1}^{m} \rho_j\, p_{k+1-j},\ k=0, 1, \ldots, \tag{5}$$

where we have implicitly adopted the convention that $p_r \equiv 0$ if $r<0$.

The probability generating function (p.g.f.) of the steady state distribution $\{p_k\}$ can be derived directly from (5), or as a special case of (4.44) of Ref. [8], p. 13.

Denoting the p.g.f. by $\pi(z)$, we obtain

$$\psi(z)=\left(1-\sum_{j=1}^{m}\rho_j\right)\Big/\left(1-\sum_{j=1}^{m}\rho_j z^j\right) \tag{6}$$

and

$$p_0=1-\sum_{j=1}^{m}\rho_j=1-\sum_{j=1}^{m}j\lambda_j/\mu=1-\rho, \tag{7}$$

where

$$\rho\equiv\sum_{j=1}^{m}\rho_j=\sum_{j=1}^{m}j\lambda_j/\mu. \tag{8}$$

Thus, ρ is the utilization factor, i.e., the ratio of the mean customer arrival rate to the mean service rate, of the multiple Poisson bulk arrival system. It follows immediately from Gaver [2, Theorem 9] that $\rho<1$ is a sufficient condition for the existence of the steady state solution and, in particular, that $p_0=1-\rho$.

That we shall be able to give explicit expressions for the p_k's is of some mathematical interest, since these expressions represent the inverse of $\psi(z)$, a function not amenable to inversion by analytical methods for arbitrary m.

4. THE EXPLICIT STEADY STATE DISTRIBUTION

The p.g.f. (6) cannot be inverted nor can the difference equation (5) be solved by algebraic methods for $m>4$. This problem turns out to be a version of the insoluble problem of determining the zeroes of at least a quintic equation. Algebraic solution for $2<m\leq 4$, while conceptually feasible, is from a practical viewpoint too cumbersome, as seems evident from Harris [5] for $m=2$. There are some indications that an analytic solution is conceptually possible through the use of the imbedded Markov chain technique and certain combinatorial probabilistic arguments, but again, for arbitrary m, this approach seems unduly tedious (see Harris [6]). Consequently, we resort to solution by inspection, which seems justified in view of the relative simplicity of the difference equation (5).

First we specify some notation. The symbol $[x]$ will denote the greatest integer less than or equal to the real number x. The multiple combinatorial symbol

$$\binom{m}{i_1,\, i_2,\, \ldots,\, i_j}=\frac{m!}{(i_1)!\ldots(i_j)!} \tag{9}$$

with

$$m=\sum_{k=1}^{j} i_k$$

will be employed with the usual conventions; viz, $0!\equiv 1$ and the left hand side of (9) will be identically 0 if $i_k<0$ for some k.

For $m=2$, i.e., bulk size limited to no more than two, it is not too unwieldly to use (5) to recursively generate several terms of $\{p_k\}$. In fact, in [7], p_k, $k=1,\ldots,8$, were generated, and by examination of these terms, we conjectured that in general

$$p_n=p_0\sum_{i=0}^{[n/2]}\rho_2^{\,i}\rho_1^{\,n-2i}\binom{n-i}{i,\, n-2i}. \tag{10}$$

It is easily demonstrated that (10) is a version of the results obtained by Harris (5) so that (10) is true for $m=2$. Proceeding similarly for $m=3$, several successive terms of $\{p_k\}$ were generated through (5) and, by inspection, we conjectured that in general

$$p_n=p_0\sum_{i_3=0}^{[n/3]}\sum_{i_2=0}^{[(n-3i_3)/2]}\rho_3^{i_3}\rho_2^{i_2}\rho_1^{n-3i_3-2i_2}\times\binom{n-2i_3-i_2}{i_3,\ i_2,\ n-3i_3-2i_2},\ k\geq 0. \tag{11}$$

At this point, it becomes apparent that these conjectures may be generalized. Specifically, a solution for arbitrary m is

$$p_{n+1}=p_0\sum_{i_m=0}^{I_m(n+1)}\sum_{i_{m-1}=0}^{I_{m-1}(n+1)}\cdots\sum_{i_2=0}^{I_2(n+1)}\left(\prod_{j=2}^{m}\rho_j^{i_j}\right)\times\binom{i_1(n+1)+\sum_{j=2}^{m}i_j}{i_m,\ \ldots,\ i_1(n+1)}\rho_1^{i_1(n+1)},\ n\geq 0, \tag{12}$$

where we define

$$i_1(n+1)=(n+1)-\sum_{j=2}^{m}ji_j, \tag{13}$$

and

$$I_k(n+1)=\left[(n+1)-\left(\sum_{j=k+1}^{m}ji_j\right)\Big/k\right]. \tag{14}$$

The brackets in (14) refer specifically to the "greatest integer in" notation. Also, (12)–(14) are written in terms of $n+1$ rather than n for later convenience.

It is easily seen that, subject to the initial condition $p_0=1-\rho$, $\rho<1$ (see (7) and (8)), any sequence $p_0, p_1, p_2, \ldots$ satisfying (5) sums to one. Thus, to show that (12) generates the explicit steady state probabilities, it suffices to show that the sequence $p_1, p_2, \ldots$ given by (12) satisfies the difference equation (5). We now proceed to demonstrate this.

We shall need the combinatorial identity (cf. Ref. [3], Ch. 24)

$$\binom{n+m}{i_1,\ \ldots,\ i_m}=\sum_{k=1}^{m}\binom{n+m-1}{i_1,\ \ldots,\ i_{k-1},\ i_k-1,\ i_{k+1},\ \ldots,\ i_m}. \tag{15}$$

Given the conventions specified following (9), (15) is well defined when $i_r=0$ for some r.

First, note that with $n=0$ in (12), the initial condition $p_1=\rho_1p_0$ is met. Then, application of (15) to (12) gives

$$p_{n+1}=\sum_{k=1}^{m}\rho_kR_k,$$

where

$$R_k=p_0\sum_{i_m=0}^{I_m(n+1)}\cdots\sum_{i_2=0}^{I_2(n+1)}\left(\prod_{\substack{j=1\\ j\neq k}}^{m}\rho_j^{i_j}\rho_k^{i_k-1}\right)\times\binom{i_1(n+1)-1+\sum_{j=2}^{m}i_j}{i_m,\ \ldots,\ i_k-1,\ \ldots,\ i_1(n+1)}, \tag{16}$$

$k=1, \ldots, m$ and $i_1\equiv i_1(n+1)$. Thus, we must show that $R_k=p_{n+1-k}$, where p_j is defined by (12).

First consider R_1, i.e., in (16) take $k=1$. Suppose that for some $r>1$, $i_r>I_r(n)$; in particular, suppose $i_r=I_r(n)+1$. Then

$$i_1(n+1)-1=n-\sum_{j=2}^{m}ji_j=n-\sum_{\substack{j=2\\ j\neq r}}^{m}ji_j-r(I_r(n)+1)=r\left\{\frac{n-\sum_{j=r+1}^{m}ji_j}{r}-\left[\frac{n-\sum_{j=r+1}^{m}ji_j}{r}\right]-1\right\}-\sum_{j=2}^{r-1}ji_j.$$

For any x, $x-[x]$ is the fractional part of x and is thus strictly less than 1; clearly, then $i_1(n+1)-1$ is negative. This causes the combinatoric in (16) to be 0 and, hence, we may replace all the upper limits therein by $I(n)$. Also in (16) with $k=1$, we can replace $i_1(n+1)-1$ by $i_1(n)$, since $i_1(n+1)-1=i_1(n)$. After making these replacements in (16), comparison with (12) indicates that $R_1=p_n$.

Now consider R_k for any $k>1$. If $i_k=0$, the combinatoric in (16) is clearly 0, so we may begin the *kth* summation with $i_k=1$. Then, after changing the index i_k to i'_k, where $i'_k=i_k-1$, and identifying $i_j=i'_j$, $j\neq 1, k$, we see that in terms of the new indices

$$(18)\qquad \begin{aligned} i_1(n+1)&=n+1-\sum_{\substack{j=2\\ j\neq k}}^{m} j i_j-k(i'_k+1)\\ &=(n-k+1)-\sum_{j=2}^{m} j i'_j=i'_1(n-k+1)\end{aligned}$$

and

$$\begin{aligned}\sum_{i=2}^{m} i_j-1&=n-k+1-\sum_{j=2}^{m} j i'_j+\sum_{j=2}^{m} i'_j\\ &=n-k+1-\sum_{j=2}^{m}(j-1)i'_j=\sum_{j=2}^{m} i'_j.\end{aligned}$$

Thus, (16) becomes for $k>1$

$$(19)\qquad R_k=p_0\sum_{i_m=0}^{I_m(n+1)}\cdots\sum_{i_k=0}^{I_k(n+1)-1}\cdots\sum_{i_2=0}^{I_2(n+1)}\left(\prod_{j=1}^{m}\rho_j^{i'_j}\right)\times\binom{i'_1(n-k+1)+\sum_{j=2}^{m} i'_j}{i'_m,\ldots,i'_1(n-k+1)}.$$

Comparison of (19) with (12) and (18) with (13) indicates that if we can replace the upper summation limits in (19) with $I_r(n-k+1)$, $r=2,\ldots,m$, then $R_k=p_{n-k+1}$ as we are attempting to prove. To this end, suppose first that $r<k$. Then

$$(20)\qquad \begin{aligned} I_r(n+1)&=\left[\frac{n+1-\sum_{\substack{j=r+1\\ j\neq k}}^{m} j i_j-k(i'_k+1)}{k}\right]\\ &=\left[\frac{n-k+1-\sum_{j=r+1}^{m} j i'_j}{k}\right]\\ &=I_r(n-k+1).\end{aligned}$$

Now suppose $r>k$ and that $i'_r>I_r(n-k+1)$ and in particular that $i'_r=I_r(n-k+1)+1$. In this case, by (18).

$$(21)\qquad i'_1(n-k+1)=n-k+1-\sum_{\substack{j=2\\ j\neq r}}^{m} j i'_j-r(I_r(n-k+1)+1)<0$$

by the same steps as in, and the argument following, (19). Thus, for $r>k$, the combinatoric in (19) is 0 for $i'_r>I_r(n-k+1)$. Finally, it is easy to see that

$$(22)\qquad I_k(n+1)-1=I_k(n-k+1).$$

By (20)–(22), each of the upper limits in (19) can be changed without effect to $I_r(n-k+1)$, $r=2, \ldots, m$.

The proof is essentially complete. We need only verify that for $k>n+1$, $R_k=0$ in keeping with our convention that $P_{n+1-k}=0$ for $k>n+1$. Now, $k>n+1$ implies $k>1$ so that (19) is appropriate. By (21), $i_1'(n-k+1)\leq n-k+1<0$ for $k>n+1$, which implies $R_k=0$.

5. QUEUE CHARACTERISTICS

For completeness and convenience we provide here some of the queue characteristics for our particular case of the $M^{(X)}M/1$ system (which we obtain from known results concerning other queueing systems).

Let N be the number of customers in the multiple Poisson bulk system with respect to the steady state distribution. Then $E(N)$ and $\mathrm{Var}(N)$ can be obtained from (5) or by the differentiation of $\psi(z)$, the p.g.f. of N (cf. Feller [1]). Specifically,

$$E(N)=\sum_{j=1}^{m} j\rho_j/p_0 \tag{23}$$

and

$$\mathrm{Var}\,(N)=\Big(\sum_{j=1}^{m} j^2\rho_j/p_0\Big)+\Big(\sum_{j=1}^{m} j\rho_j/p_0\Big)^2, \tag{24}$$

where p_0 can be expressed in terms of the system parameters as in (7), and is itself an important queue characteristic.

Other characteristics of importance in any bulk arrival process are waiting times for groups and waiting times for individual customers. When a bulk is a physical unit, as in the plywood example, waiting times for individual customers may not be of practical importance, for a customer cannot, or does not, leave the system until the bulk of which he is a member leaves. In many situations, however, customer waiting times will have some significance. Thus, when families eat in restaurants, the time which the entire family waits to be served is the significant factor, but when long distance buses stop for meals, individual passengers are generally more concerned with their own, rather than the group's, waiting time.

Let W_g be the wait on queue for a bulk when the system is in equilibrium condition. That is W_g is the elaspsed time from the arrival of the bulk to the start of service to its first member. We can calculate $E(W_g)$ for the present model from e.g. (7.17) of Gaver [2] or by suitably modifying. well-known results from M/G/1. We obtain.

$$E(W_g)=\rho[2\mu(1-\rho)]^{-1}\Big(1+\sum_{j=1}^{m} j^2\lambda_j\Big/\sum_{j=1}^{m} j\lambda_j\Big). \tag{25}$$

Now let T_g be the total time a bulk spends in the system and let X refer, as before, to bulk size. Then for a bulk of specified size, e.g. J, the expected total waiting time is $E(T_g|X=J)=E(W_g)+J/\mu$ and, in general, $E(T_g)=E(W_g)+\rho/\lambda$.

Let T_c denote the time an individual customer spends on queue and let A_j be the event that a customer is the jth served within a bulk. Assume that service within a bulk is random, i.e., $P(A_j|X=J)=1/J$, $1\leq j\leq J$ and 0 otherwise. Then, if bulk size is specified, e.g. J, the expected time on queue for a member of that bulk is

$$E(T_c|X=J)=\sum_{j=1}^{J} E(T_c|A_j, X=J)P(A_j|X=J) \tag{26}$$

$$=E(W_g)+\sum_{j=1}^{J} [(j-1)/\mu](1/J)$$

$$=E(W_g)+(1/2\mu)(J-1).$$

For the general case in which bulk size is not specified, the expected customer time on queue is $E(T_c)$, where

$$E(T_c)=\sum_{J=1}^{m} E(T_c|X=J)P(X=J)$$

$$=E(W_g)+(1/2(\rho/\lambda-1/\mu)).$$

Expected total time in system rather than on queue can be obtained from (26) and (27) by adding $1/\mu$ to $E(T_c)$ or $E(T_c|X=J)$, whichever is appropriate.

It is of some interest to compare the multiple Poisson bulk arrival system with the simple, i.e., the $M/M/1$, queueing system, where we assume that the customer mean arrival and service rates are the same for both systems. Let $N_s(N)$ denote the number of customers in the $M/M/1$ (multiple Poisson bulk arrival) system. That $E(N)-E(N_s)>0$ for the general $M^{(X)}/G/1$ bulk arrival system was pointed out by Gaver [2]. In particular, for the present model, by (23) and well-known results for $M/M/1$,

$$E(N)-E(N_s)=\sum_{j=1}^{m} j(j-1)\lambda_j/[2(1-\rho)\mu]$$

or equivalently

$$E(N)/E(N_s)=\sum_{j=1}^{m} j(j+1)\lambda_j\Big/\Big(2\sum_{j=1}^{m} j\lambda_j\Big).$$

In fact, in our computer simulations of the multiple Poisson bulk arrival system, we always observed that N is stochastically greater than N_s, i.e., $P(N>n)\geq P(N_s>n)$, for all n (cf. Ref. [7], Tables III–IV). A final comparison with $M/M/1$ is that while $P(N_s=n)$ decreases monotonically in n, $P(N=n)$ does not necessarily (cf. Ref. [7], Fig. 7).

REFERENCES

[1] Feller, W., *An Introduction to Probability Theory and Its Applications*, Vol. 1, Third Ed., (John Wiley and Sons, Inc., New York 1966).

[2] Gaver, D. P., Jr., "Imbedded Markov Chain Analysis of a Waiting Line Process in Continuous Time," Annals of Mathematical Statistics, *30*, 698–720 (1959).

[3] Goldberg, K., M. Newman, and E. Haynsworth, "Combinatorial Analysis" in M. Abramowitz and I. A. Stegun, Eds., *Handbook of Mathematical Functions*, U.S. Government Printing Office, Washington, D.C., (1964).

[4] Gross, D., and C. M. Harris, *Fundamentals of Queueing Theory*, (John Wiley and Sons, Inc., New York 1974).

[5] Harris, C. M., "Queues with Multiple Poisson Input," Journal of Industrial Engineering, *17*, 454–60 (1966).

[6] Harris, C. M., "Some Results for Bulk-Arrival Queues with State Dependent Service Times," Management Science, *16*, 313–26 (1970).

[7] Jensen, G. L., "The Multiple Bulk Poisson Arrival Queueing System," Unpublished Ph.D. dissertation, University of Tennessee, (1973).

[8] Kleinrock, L., *Queueing Systems, Vol. I. Theory*, (John Wiley and Sons, Inc., New York 1975).

[9] Ross, S. L., *Introduction to Probability Models*, (Academic Press, New York 1972).

A PROCEDURE FOR GENERATING TIME-DEPENDENT ARRIVALS FOR QUEUEING SIMULATIONS*

George S. Fishman

University of North Carolina at Chapel Hill
Chapel Hill, N.C.

Edward P. C. Kao

University of Houston
Houston, Texas

ABSTRACT

This paper presents a method for modeling cyclic inputs to a congested system in a discrete event digital simulation. Specifically, we express the mean of the interarrival time conditional on the last arrival as a linear combination of harmonic components whose coefficients can be determined by stepwise regression. We also assume that the conditional interarrival time normalized by its corresponding mean follows a distribution that is independent of time. The result can, in turn, be used to generate the desired input for a simulation, An example based on a set of actual data is used to illustrate the process of parameter estimation for the model.

1. INTRODUCTION

Although statistical methods play a central role in the digital simulation of queueing systems, one area in which these methods have been notably under-utilized is the construction of arrival generators for systems whose arrival frequencies vary with time. For example, it is not unusual to find arrival patterns that vary with time of day and day of week. Failure to acknowledge these patterns, when they exist, can lead to serious distortions in system behavior. In particular, treating interarrival times as independent and identically distributed when they, in fact, exhibit a cyclic pattern removes the clustering of arrivals associated with successively shorter interarrival times. Since this clustering is a principal source of congestion, the effect of overlooking dependence is to distort the congestion pattern.

To put the problem in perspective, we shall first describe the principal issues that arise in attempts to characterize time-dependent arrivals by *empirical* distributions. For expository purposes, we consider an arrival pattern that depends on the day of week but, within day interarrival times, is independent and uniformly distributed. Let $\{X_{jk}; j=1, \ldots, 7; k=1, \ldots, K\}$ be a sample record of the number of daily arrivals over K weeks. Then

*This work was partially supported under an Office of Naval Research Contract N0014-67-A-0097-0028-119).

$$F_j(i)=K^{-1}\sum_{k=1}^{K} I_{(X_{jk},\,\infty)}(i) \qquad i=0, 1, \ldots, \infty,$$

where

$$I_{(X_{jk},\,\infty)}(i)=\begin{cases}0, & i<X_{jk}\\ 1, & i\geq X_{jk}\end{cases}$$

denotes the sample cumulative distribution function for the number of arrivals on day j of each week. Suppose that we want to generate an arrival on day j'. Let U denote a uniform deviate on (0,1). Then the number of arrivals on day j' is $I_{j'}=\inf\,[i{:}U\leq F_j(i);\ j=j' \pmod 7]$, and the k^{th} arrival on that day occurs at $j'+V_k$, where $\{V_k;\ k=1, \ldots, I_{j'}\}$ is a sequence of independent uniform deviates. The extension of the above approach to include other cyclic variations is straightforward (e.g., see Ref. [4], p. 6). The appeal of this approach lies in its simplicity and complete reliance on untransformed uniform deviates. Although alternative methods exist for generating arrival times from empirical distributions, the above procedure is common enough so that the statistical and computational issues it raises merit serious attention.

Statistical Issues.

Since empirical distributions only represent realizations of random phenomena, discontinuities in frequency distributions and presence or absence of extreme values are to be expected. This important drawback is intrinsic to the use of statistical data directly as input. A large sample could only partially rectify the situation. Also, splitting a complete cycle into segments with homogeneous arrival rates could cause abrupt behavioral changes between adjacent segments.

Computational Issues

Since the arrival times generated in the prescribed manner are unordered within a day, the simulation must devote time to ordering them in the process of scheduling each arrival. This leads to multiple arrival notices in the list of scheduled events, a situation that reduces the computational efficiency of the simulation. An alternative approach is to generate all arrivals in a preprocessing program that orders them chronologically. However, the resulting saving in computational efforts is not at all clear.

In the next section, we shall introduce a procedure that uses interarrival times to generate time-dependent arrivals. Although the procedure does not totally resolve these issues, they at least reduce their saliency, especially the statistical aspects of the problem. Finally, we shall present an example based on actual data to illustrate the process of parameter estimation for the model.

2. THE MODEL

Background

In this section, we describe a procedure that uses interarrival time data to construct an arrival generator for queuing simulations. Let T_i denote the arrival time of the i^{th} arrival, and $S_i=T_i-T_{i-1}$ denote the interarrival time between the $(i-1)^{\text{st}}$ and i^{th} arrivals. Suppose that we represent the i^{th} interarrival time S_i conditional on T_{i-1} as

$$S_t = \lambda(T_{t-1}) + \epsilon_t, \tag{1a}$$

$$\lambda(T_{t-1}) = E(S_t | T_{t-1}) = a_0 + \sum_{j=1}^{n} (a_j \cos \theta_j T_{t-1} + b_j \sin \theta_j T_{t-1}), \tag{1b}$$

where we assume that $\lambda(T_{t-1}) > 0$ and $\{\epsilon_t | T_{t-1}\}$ is a sequence of independent random variables with zero mean and domain on the half open interval $[-\lambda(T_{t-1}), \infty)$.* In using the harmonic function to represent the mean conditional interarrival time, the choice of the $\{\theta_j\}$ determines a set of frequencies that account for cyclic behavior in an arrival pattern. As an example, if we want to take the within-week cyclic variations into consideration (e.g., the variations due to hour of the day and day of the week), and times are measured in hours, then we would use $\theta_1 = \pi/168$ and $\theta_j = j\theta_1$ for $j = 2, \ldots, 168$. The $\{a_j\}$ and $\{b_j\}$ in (1b) remain to be estimated from the data.

To facilitate the generation of the S_t during a simulation, we define a normalized interarrival time R_t conditional on T_{t-1} as

$$R_t = S_t / \lambda(T_{t-1}) = 1 + \epsilon_t / \lambda(T_{t-1}). \tag{2}$$

We assume that $\{R_t | T_{t-1}\}$ is a sequence of independent, identically distributed nonnegative random variables.* We shall have more to say about R_t in the sequel.

Parameter Estimation

The estimation of the $\{a_j\}$ and $\{b_j\}$ in (1b) can be done through the use of some form of regression analysis, provided that (a) $\lambda(T_{t-1}) > 0$, and $\epsilon_t > -\lambda(T_{t-1})$. Imposing the two constraints on a regression analysis creates difficulties with regard to the sampling properties of the method. However, for all practical purposes we propose to circumvent the complication by performing an unconstrained stepwise regression to determine which of the $2n+1$ coefficients contribute significantly to $\lambda(T_{t-1})$ and then examining the results to determine whether the constraints have been violated.

Before we postulate an underlying distribution for the normalized interarrival time R_t conditional on T_{t-1}, we shall first look at its sample mean $E(\hat{R}_t | T_{t-1})$. Let $\hat{\lambda}(T_{t-1})$ be the estimate of $\lambda(T_{t-1})$ from the stepwise regression and consider the sequence $\{\hat{R}_t | T_{t-1}\}$, where

$$\begin{aligned}
\hat{R}_t = \frac{S_t}{\hat{\lambda}(T_{t-1})} &= \frac{S_t}{\lambda(T_{t-1}) + [\hat{\lambda}(T_{t-1}) - \lambda(T_{t-1})]} \\
&= \frac{S_t/\lambda(T_{t-1})}{1 - \{[\lambda(T_{t-1}) - \hat{\lambda}(T_{t-1})]/\lambda(T_{t-1})\}} \\
&= (S_t/\lambda(T_{t-1}))\left[1 + \sum_{j=1}^{\infty} \{[\lambda(T_{t-1}) - \hat{\lambda}(T_{t-1})]/\lambda(T_{t-1})\}^j\right].
\end{aligned}$$

Taking expectations of the above expression, we obtain

$$E(\hat{R}_t | T_{t-1}) = 1 + (1/\lambda(T_{t-1}))\, E\left(S_t \sum_{j=1}^{\infty} \{[\lambda(T_{t-1}) - \hat{\lambda}(T_{t-1})]/\lambda(T_{t-1})\}^j\right). \tag{3}$$

*These assumptions can easily be verified in a regression analysis. $\{Y_t | X\}$ denotes a sequence of random variables conditional on X.

If the sample size is large, we would expect the deviations $\lambda(T_{i-1})-\lambda(T_{i-1})$ to be small relative to $\lambda(T_{i-1})$. Consequently, we choose to ignore the rightmost term of (3). In practice, a quick way to check the relative importance of the term is to compute $\Sigma R_i/N$ (N is the sample size) and check if it close to one. Henceforth, our model specification implies that R_i is positive with mean unity and identically distributed for all i.

To make a distributional assumption for R_i, at least three distributions deserve consideration. They are the gamma, lognormal, and Weibull distributions. Imposing the constraint that R_i has a mean unity reduces each of these to a one-parameter distribution. The selection of a proper underlying distribution among the three requires a consideration of the "goodness-of-fit" of the data under the postulated distributions as well as the relative computational efficiency in generating random variables under each choice. A detailed discussion of the issues involved can be found in Ref. [2], pp. 208–213 and Ref. [4], pp. 10–12. In the example to be given in the next section, we find that the use of the Weibull distribution is adequate as far as the prescribed issues are concerned.

For the Weibull case, R_i has the following probability density function:

$$f_{R_i}(r)=\begin{cases}\alpha[\Gamma(1/\alpha+1)]^{\alpha}r^{\alpha-1}e^{-[r\Gamma(1/\alpha+1)]^{\alpha}}, & 0\leq r\leq\infty\\ 0, & \text{elsewhere.}\end{cases} \tag{4}$$

The maximum likelihood estimation of α is the solution to the equation

$$N/\hat{\alpha}+N\ln\Gamma(1/\hat{\alpha}+1)-N\Psi(1/\hat{\alpha}+1)/\hat{\alpha}+\Sigma\ln R_i-[\Gamma(1/\hat{\alpha}+1)]^{\hat{\alpha}}\Sigma R_i^{\hat{\alpha}}\ln R_i$$
$$-[\Gamma(1/\hat{\alpha}+1)]^{\hat{\alpha}}[\ln\Gamma(1/\hat{\alpha}+1)-\Psi(1/\hat{\alpha}+1)/\alpha]\Sigma R_i^{\hat{\alpha}}=0,$$

where $\psi(\cdot)$ is the digamma function. When $\Sigma\hat{R}_i/N$ is very close to unity, the unconstrained maximum likelihood procedure described in Ref. [2] (i.e., a procedure without presuming that R_i has an unit mean) provides a convenient alternative for estimating α.

3. AN EXAMPLE

To illustrate the use of the procedure presented in the last section, we use a set of sequences of successive patient arrivals to a coronary care unit in a hospital in New Haven, Connecticut.* The sample record covers a two-year period during which arrivals not admitted to the unit were not recorded. However, we were able to identify the sequences that corresponded to the time segments in which successive arrivals were all admissions. Based on these data, an unconstrained stepwise multiple regression with $n=168$ yields an F value of 5.257 with 18 and 223 degrees of freedom, which is significant at the 99.9 percent level. The fact that $R^2=0.236$ (i.e., the proportion of variations attributable to the regression equation) should by no means be construed as less than satisfactory. A priori, we know that a large component of random variation exists in the data. Fitting the harmonic terms removes the time-dependent share of variations. To check for correlation between successive normalized interarrival times, we analyzed the pairs (R_i, R_{i+1}) obtained from all sequences containing at least three successive arrivals and found no evidences to reject their stochastic independence.

*The significance of fitting these data to a time series model for simulation purposes is described in Ref. [3].

Having found $\Sigma R_t/N=0.9984$, we decided to use the one-parameter Weibull given in (4) as the underlying distribution for R_t to account for time-independent stochastic behavior. The unconstrained maximum likelihood procedure of Ref. [1] gave $\hat{\alpha}=1.413$. To evaluate the "goodness-of-fit", we compared the sample cumulative distribution of the R_t with the estimated cumulative Weibull distribution $1-e^{-r^{\hat{\alpha}}}$ and found that the fit was unusually close in the tails but not as close in the central range. For details, see Ref. [4].

As mentioned earlier, the regression analysis used was unconstrained. A careful check of $\hat{\lambda}(T)$ for $0<T\leq 168$ revealed that one negative value did indeed occur. Although negative values seem inconsistent with a model whose parameter estimation relies on positive data, we offer the following rationalization. In fitting a harmonic function, a least-squares fit may produce negative values when data contain rapid downward shifts in mean values over relatively short intervals. That is, the minimizing unconstrained curve must be regarded as an interpolation between sample data points, and it unfortunately can produce negative values when interpolating between the last point in a rapidly decreasing sequence of interarrival times and the first point in a rapidly increasing sequence of interarrival times.

One way to handle the problem of negative values is to proceed with the use of the estimated $\hat{\lambda}(T)$, generate all interarrival times prior to the simulation, and then order them chronologically. Although this approach is computational time consuming, it solves the problem and overcomes most of the inadequacies of the empirical approach mentioned in Section 1. An alternative approach that guarantees $\hat{\lambda}(T)>0$ is suggested in Ref. [4]. The procedure involves working with the logarithm of S_t in (1a). Unfortunately, when it was applied to the same set of data, we found that the fit of the underlying distribution for R_t was considerably poorer than that of the first approach (Ref. [4], pp. 20–22).

4. GENERATION OF INTERARRIVAL TIMES

In generating interarrival times S_t using the aforementioned procedure, we first recall that $S_t=R_t\hat{\lambda}(T_{t-1})$. If the last arrival occurred at time T_{t-1} and the normalized conditional interarrival time R_t follows the Weibull distribution, then the interarrival time for the next arrival S_t is given by

$$S_t=\frac{\hat{\lambda}(T_{t-1})}{\Gamma(1/\hat{\alpha}+1)}\,(-\ln U)^{1/\hat{\alpha}},$$

where U is to be generated from the uniform distribution $\mu(0,\,1)$ (e.g., see Ref. [2], p. 211). To save computation time, we can table the values of $\hat{\lambda}(t)$ for suitably discretized values of t in a complete cycle (e.g., a week) before carrying out the simulation runs.

REFERENCES

[1] Cohen, C., "Maximum Likelihood Estimation in the Weibull Distribution Based on Complete and on Censored Samples," Technometrics, 7, 579–588 (1965).

[2] Fishman, G. S., *Concepts and Methods in Discrete Event Digital Simulation* (Wiley, New York, 1973).

[3] Fishman, G. S., and E. P. C. Kao, "The Analysis of Hospital Arrival Patterns for Heart Attack Patients," Health Services Research Training Program, Institution for Social and Policy Studies, Yale University, (1973).

[4] Fishman, G. S., and E. P. C. Kao, "Arrival Generators for Queueing Simulations," Technical Report 28–5, Department of Administrative Sciences, Yale University (July 1974).

CONFIDENCE INTERVALS IN DISCRETE EVENT SIMULATION: A COMPARISON OF REPLICATION AND BATCH MEANS*

Averill M. Law

University of Wisconsin
Madison, Wisconsin

ABSTRACT

Suppose that we have enough computer time to make n observations of a stochastic process by means of simulation and would like to construct a confidence interval for the steady-state mean. We can make k independent runs of m observations each ($n=k\cdot m$) or, alternatively, one run of n observations which we then divide into k batches of length m. These methods are known as replication and batch means, respectively. In this paper, using the probability of coverage and the half length of a confidence interval as criteria for comparison, we empirically show that batch means is superior to replication, but that neither method works well if n is too small. We also show that if m is chosen too small for replication, then the coverage may decrease dramatically as the total sample size n is increased.

1. INTRODUCTION

Let $\{X_i,\ i\geqslant 1\}$ be a stochastic process for which we would like to estimate the steady-state mean μ:

$$\mu=\lim_{n\to\infty}\sum_{i=1}^{n}X_i/n \text{ (with probability 1)}$$

$$=\lim_{n\to\infty}E\sum_{i=1}^{n}X_i/n.$$

(These limits exist and are equal for all processes considered in this paper; see Ross [12], p. 98.) Given enough computer time to make n observations of the process by means of simulation, how should one construct a confidence interval (c.i.) for μ? The difficulty is that for most simulations, the observed process is nonstationary and autocorrelated. Thus, the techniques of classical statistical analysis for independent identically distributed (i.i.d.) observations are not directly applicable.

Five methods have been suggested in the simulation literature for solving the above problem: replication, batch means, spectrum analysis, autoregressive representation, and regeneration

*This research was supported by the Office of Naval Research under contract N00014-76-C-0403 (NR 047-145).

cycles. (See Crane and Iglehart [3, 4], Fishman [6], and Iglehart [10] for descriptions of these methods.) However, there has been no definitive effort made to determine which of the methods is the best for a given simulation situation. Thus, a simulator who is actually interested in estimating a steady-state mean may not know which method to employ.

In this paper, which is the first in a series on confidence intervals in simulation, we analyze and compare the widely used methods of replication and batch means. The methods are similar in philosophy in that both try to avoid autocorrelation by breaking the data into "independent" segments. The sample mean of the data in each segment is computed and the analysis for i.i.d. observations is applied to the sample means to construct a c.i. for the steady-state mean. Using coverage and half length as criteria for comparison, we conclude from simulations of several queueing and inventory systems that batch means is superior to replication, but that neither method works well if the total sample size n is too small.

The remainder of this paper is organized as follows. Section 2 describes the two methods in detail. The methods are empirically compared in Section 3. Section 4 explains why the empirical results differ from theory and Section 5 offers suggestions as to how the methods may be improved.

2. DESCRIPTION OF THE TWO METHODS

A. Replication

Suppose we make k independent simulation runs, each of length m observations ($n=k\cdot m$). We accomplish the independence of runs by starting each run from scratch (for queueing simulations this usually means that no customers are present at time zero) and by using a different stream of random numbers for each run. Let $\overline{X}_i(m)$ ($i=1, 2, \ldots, k$) be the sample mean of the m observations in the ith run. The $\overline{X}_i(m)$'s are i.i.d. random variables (r.v.'s) since the runs themselves are. We use

$$\overline{\overline{X}}(k, m)=\sum_{i=1}^{k} \overline{X}_i(m)/k$$

as our point estimator of the steady-state mean μ.

Let $\mu(m)=E[\overline{X}_i(m)]$. Then by the definition of μ, $\mu(m)\rightarrow\mu$ as $m\rightarrow\infty$. Furthermore, if $\mu(m)=\mu$ and $0<\sigma^2[\overline{X}_i(m)]<\infty$, then by the classical central limit theorem we have

$$(1)\qquad \frac{\overline{\overline{X}}(k, m)-\mu}{\sqrt{\sigma^2[\overline{\overline{X}}(k, m)]}} \xrightarrow{\mathfrak{D}} N(0, 1) \text{ as } k\rightarrow\infty,$$

where $N(0, 1)$ is a mean 0, variance 1, normal r.v., and $\xrightarrow{\mathfrak{D}}$ denotes convergence in distribution. Furthermore, (1) remains true if $\sigma^2[\overline{\overline{X}}(k, m)]$ is replaced by

$$(2)\qquad \hat{\sigma}^2[\overline{\overline{X}}(k, m)]=\sum_{i=1}^{k}[\overline{X}_i(m)-\overline{\overline{X}}(k, m)]^2/k(k-1).$$

If the $\overline{X}_i(m)$'s are normally distributed, then the ratio

$$[\overline{\overline{X}}(k, m)-\mu]/\sqrt{\hat{\sigma}^2[\overline{\overline{X}}(k, m)]}$$

has the t distribution with $k-1$ degrees of freedom (d.f.), and an exact $100(1-\alpha)\%$ c.i. for μ is given by

$$\bar{\bar{X}}(k, m) \pm t_{k-1,\, 1-\alpha/2}\sqrt{\hat{\sigma}^2[\bar{\bar{X}}(k, m)]}, \tag{3}$$

where $t_{k-1,\, 1-\alpha/2}$ is the $1-\alpha/2$ point for a t distribution with $k-1$ d.f. Even if the $\bar{X}_i(m)$'s are not normally distributed, it is common practice when k is small to use (3) to construct a c.i. for μ.

There are two potential sources of error when using replication to construct a c.i. for a steady-state mean: the fact that $\mu(m) \neq \mu$ and the nonnormality of the $\bar{X}_i(m)$'s. These errors will be discussed in Section 4.

B. Batch Means

One disadvantage of replication is that $E[\bar{\bar{X}}(k, m)] = \mu(m) \neq \mu$ for any k, so that for fixed m, $\bar{\bar{X}}(k, m)$ is a biased estimator of μ no matter how many replications are made. Since $\mu(m) \rightarrow \mu$ as $m \rightarrow \infty$, suppose, as an alternative to replication, we now make one long run of length n and then divide the resulting observations $X_1, X_2, \ldots, X_n$ into k batches each of length m. Let $\bar{X}_i(m)$ $(i=1, 2, \ldots, k)$ be the sample mean of the m observations in the ith batch. Once again we use $\bar{\bar{X}}(k, m)$ as our point estimator of μ.

If we choose m sufficiently large, then the $\bar{X}_i(m)$'s will be essentially uncorrelated (see Tables 5 and 7), and we can estimate $\sigma^2[\bar{\bar{X}}(k, m)]$ by (2). If m is large enough so that the $\bar{X}_i(m)$'s are approximately normally distributed in addition to being uncorrelated, then the $\bar{X}_i(m)$'s are also "independent" (see Ref. [6], p. 142). We thus have essentially the same situation as for replication and (3) may be used to construct a c.i. for μ.

There are three potential sources of error when using batch means to construct a c.i. for a steady-state mean: the correlation between the $\bar{X}_i(m)$'s, the fact that the $\bar{X}_i(m)$'s are not identically distributed with mean μ, and the nonnormality of the $\bar{X}_i(m)$'s.

For expository convenience, we will henceforth use the phrase "point estimator bias" to mean $\mu(m) \neq \mu$ for replication, and to mean the $\bar{X}_i(m)$'s are not identically distributed with mean μ for batch means.

3. EMPIRICAL COMPARISON

In order to compare replication and batch means, we simulated several well-known queueing and inventory systems for which analytical results are available. The results of these simulations are presented in this section.

The random numbers $\{U_i, i \geq 1\}$ used in this paper were generated from the following generator which is available on our version of the Univac 1110:

$$Y_i = (5^{15}Y_{i-1} + 1) \bmod 2^{35} \quad (i=1, 2, \ldots)$$

$$U_i = Y_i/2^{35} \quad (i=1, 2, \ldots),$$

where Y_0 is a given seed. For a discussion of this generator, see Coveyou and Macpherson [2].

A. *M*/*M*/1 Queue

The first stochastic system we considered was the *M*/*M*/1 queue. We let $E(A)=1$ (the mean interarrival time), $E(S)=0.9$ (the mean service time), $\rho=E(S)/E(A)=0.9$ (the traffic intensity), and then simulated the stochastic process $\{D_i, i\geq 1\}$, where D_i is the delay in queue (not including service time) of the *i*th customer and, for our study, $D_1=0$ (i.e., no customers are present at time zero). Our objective was to construct 90% c.i.'s for the steady-state mean delay in queue, $d=8.1$. We performed 400 independent simulation experiments; for each experiment, we considered $n=1600$, 3200, 6400, 12800 and $k=5$, 10, 20, 40; for each n and k we constructed a c.i. for d using both replication and batch means. However, the longer runs were continuations of the shorter runs, so results for different values of n are not independent. Furthermore, for a given value of n, exactly the same stream of random numbers was used to compute the delays for the two methods.

In Figure 1 we plot, for each n and k and each method, the average half length of the 400 c.i.'s vs the proportion of the 400 c.i.'s which covered d. The same proportions of coverage are also given in Table 1. By way of example, for replication with $n=1600$ and $k=5$ (i.e., 5 runs of length 320), 211 out of the 400 (or 52.75%) 90% c.i.'s contained d and the average half length was 3.11. From Figure 1 we can see that for a given value of n, the curve for replication lies above that for batch means. This implies that for a specified half length, batch means has greater coverage than replication or, alternatively, for a specified coverage, batch means has a shorter half length than replication.

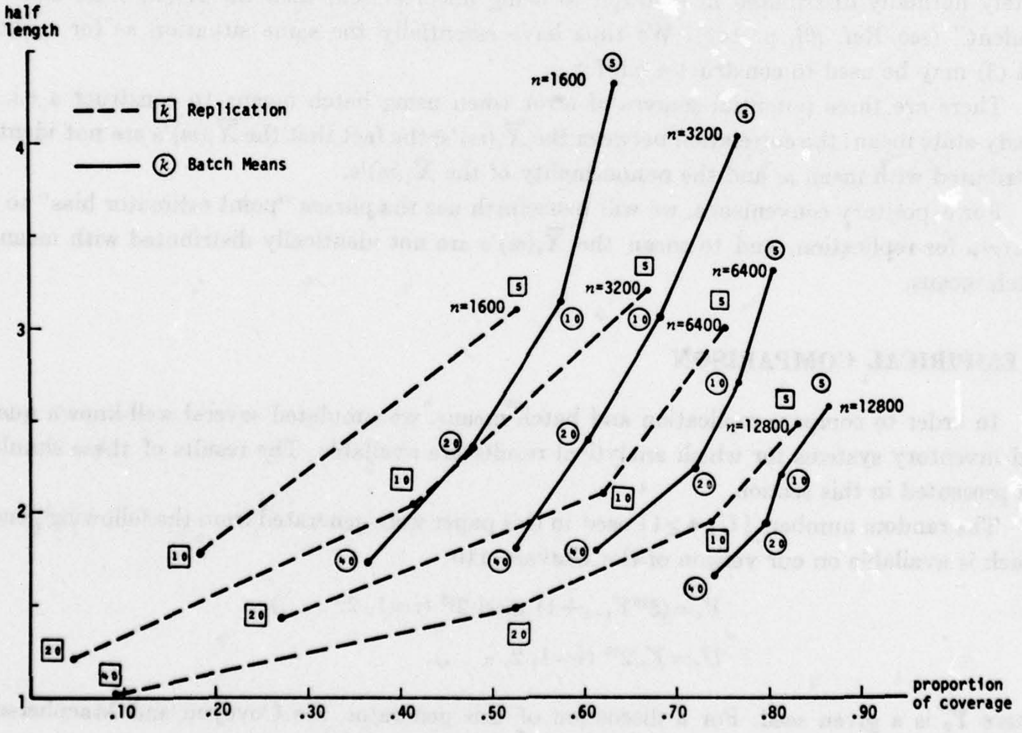


FIGURE 1. Average half length vs proportion of coverage for various 90% confidence intervals for $d=8.1$ in an *M*/*M*/1 queue with $\rho=0.9$.

TABLE 1. *Proportion of Coverage of Various 90% Confidence Intervals for d=8.1 in an M/M/1 Queue with $\rho=0.9$.*

	Replication				Batch Means			
n \ k	5	10	20	40	5	10	20	40
1600	0. 5275	0. 1825	0	0	0. 6325	0. 5750	0. 4750	0. 3650
3200	0. 6700	0. 4175	0. 0450	0	0. 7675	0. 6825	0. 6050	0. 5225
6400	0. 7525	0. 6200	0. 2700	0	0. 8050	0. 7675	0. 7200	0. 6125
12800	0. 8325	0. 7375	0. 5275	0. 0900	0. 8650	0. 8125	0. 7975	0. 7400

Table 1 shows that batch means always has greater coverage than replication. To see whether these observed differences were statistically significant, we performed a paired-t test for each n and k. Each test was based on a sequence of 400 pairs of numbers, with each number in each pair being a 1 or 0, depending on whether or not the c.i. covered d. For 15 out of 16 tests, the differences were significant at least at the 99% level; in the other case, the observed level of significance was 95%.

Notice from Table 1 that if m is held fixed and k (and thus n) is increased (perhaps in an effort to get a shorter c.i.), then the coverage for replication may be drastically reduced. (See, for example, the coverages on the main diagonal.) This degradation occurs because $E[\overline{\overline{X}}(k, m)]=\mu(m)$ for all k, but $\sigma^2[\overline{\overline{X}}(k, m)]=\sigma^2[\overline{X}_i(m)]/k$. Thus, as k is increased, a shorter c.i. is constructed around $\mu(m)\neq\mu$, resulting in a decrease in coverage.

To determine the generality of the above results, we also simulated the M/M/1 queue with $\rho=0.5$ and 0.7, the M/M/2 queue with $\rho=0.9$, and the M/M/1/M/1 queue (the output of an M/M/1 queue is the input to another single-server queue with exponential service) with $\rho=0.9$ for each server. In each case, we obtained results similar to those above; in particular, we found batch means superior to replication.

B. (s, S) Inventory System

The second type of stochastic system we considered was an (s, S) inventory system with zero delivery lag and backlogging. Let X_i, Y_i, and Q_i denote, respectively, the amount of inventory on hand before ordering, the amount of inventory on hand after ordering, and the demand, each in period i. If $X_i<s$, then we order $S-X_i$ items ($Y_i=S$) and incur an ordering cost $K+c\cdot(S-X_i)$. If $X_i\geq s$, then no order is placed ($Y_i=X_i$) and no ordering cost is incurred. After Y_i has been determined, then the demand Q_i occurs. If $Y_i-Q_i\geq 0$, then a holding cost $h\cdot(Y_i-Q_i)$ is incurred. If $Y_i-Q_i<0$, then a shortage cost $\pi\cdot(Q_i-Y_i)$ is incurred. In either case, $X_{i+1}=Y_i-Q_i$. For further discussion of this inventory system see Wagner [13], p. A19.

For our study, we let Q_i be a Poisson r.v. with mean 25, $s=17$, and $S=57$ (these values, which are approximately optimal, were computed from a normal approximation given in [13], p. A40), $X_1=S$, $K=32$, $c=3$, $h=1$, $\pi=5$, and then simulated the stochastic process $\{E_i, i\geq 1\}$, where E_i is the total cost (expenditure) in period i. Our objective was to construct 90% c.i.'s for the steady-state mean cost per period, $e=112.108$. Once again, we performed 400 independent experiments and let $k=5$, 10, 20, 40; however, now we let $n=320$, 640, 1280, 2560. In Tables 2 and 3 we give the proportion of the 400 c.i.'s which covered e and their average half length, respectively. It is not possible to give a meaningful graph of average half length vs proportion of coverage since the coverages for batch means are all over 0.90. (See Subsection 4.B for an explanation.) However, it is clear from Table 2 that in terms of coverage, batch means is far superior to replication; in fact, the observed differences are significant at least at the 99% level for each n and k. We feel that for most purposes, these large increases in coverage more than justify the small increases in half length observed in Table 3.

4. ANALYSIS OF ERRORS

We have seen that the actual coverages of c.i.'s produced by batch means and replication may be considerably different from those desired. In the following three subsections, we use the

TABLE 2. *Proportion of Coverage of Various 90% Confidence Intervals for $e=112.108$ in an (s, S) Inventory System with $s=17$, $S=57$, and Poisson Demand.*

	Replication				Batch Means			
k / n	5	10	20	40	5	10	20	40
320	0.4625	0.0175	0	0	0.9375	0.9575	0.9825	0.9925
640	0.6850	0.1525	0	0	0.9325	0.9425	0.9750	0.9875
1280	0.7925	0.3950	0.0025	0	0.9200	0.9350	0.9525	0.9750
2560	0.8500	0.5925	0.0875	0	0.9175	0.9425	0.9425	0.9600

TABLE 3. *Average Half Length of Various 90% Confidence Intervals for $e=112.108$ in an (s, S) Inventory System with $s=17$, $S=57$, and Poisson Demand.*

	Replication				Batch Means			
k / n	5	10	20	40	5	10	20	40
320	2.018	1.882	1.822	1.729	2.377	2.281	2.447	2.630
640	1.388	1.264	1.254	1.263	1.527	1.419	1.486	1.639
1280	0.950	0.863	0.855	0.873	0.998	0.924	0.943	1.014
2560	0.652	0.592	0.576	0.588	0.667	0.613	0.611	0.642

$M/M/1$ queue to determine how the potential sources of error mentioned in Section 2 affect the two methods. We will explicitly mention the above inventory system only when a source of error affects it differently from the manner in which it affects the $M/M/1$ queue. Our ultimate objective is, of course, to learn how best to employ the two methods.

A. Point Estimator Bias

A requirement of both methods is that the $\overline{X}_i(m)$'s be identically distributed with mean μ. However, for most real-world simulations this requirement is not met, and causes a difficulty in estimating μ which we called point estimator bias. Let $\overline{D}(l)$ be the sample mean of $D_1, D_2, \ldots, D_l$. We can graphically see the point estimator bias for the $M/M/1$ queue in Figure 2 where we plot $E[\overline{D}(l)|D_1=0]$ (see Heathcote and Winer [9]) as a function of l. Note that $E[\overline{D}(l)|D_1=0]\neq d$ for any l, but approaches it as l increases (as it should by the definition of d).

To determine the degradation in coverage caused by the point estimator bias, we simulated the stationary $M/M/1$ queue; that is, an $M/M/1$ queue where the number of customers found by the "first" arrival is a r.v. with the stationary number in system distribution (see Gross and Harris 8], p. 47). (In this case, it is easy to show that D_1 has the stationary delay in queue distribution.) Thus, $\{D_i, i\geq 1\}$ is a stationary stochastic process, $E[\overline{D}(l)]=d$ for each l, and there is no point estimator bias. We performed 400 independent simulation experiments using the same values of n and k and the same stream of random numbers as we did for the usual $M/M/1$ queue. Table 4 gives, for each n and k and each method, the proportion of the 400 c.i.'s which contained d. A comparison of Tables 1 and 4 shows that coverage in the stationary case is greatly improved for replication but only slightly improved for batch means. This is not surprising since, for example, $E[\overline{D}(320)|D_1=0]=6.01$ but $E[\overline{D}(1600)|D_1=0]=7.59$ (see Figure 2). We conclude that for the values of n considered, point estimator bias is a major source of degradation for replication but has little effect on batch means.

TABLE 4. *Proportion of Coverage of Various 90% Confidence Intervals for $d=8.1$ in a Stationary $M/M/1$ Queue with $\rho=0.9$.*

	Replication				Batch Means			
k \ n	5	10	20	40	5	10	20	40
1600	0. 7800	0. 8100	0. 8525	0. 8925	0. 6875	0. 6300	0. 5350	0. 4050
3200	0. 8150	0. 8200	0. 8450	0. 8975	0. 7925	0. 7075	0. 6275	0. 5300
6400	0. 8450	0. 8350	0. 8550	0. 8900	0. 8275	0. 7950	0. 7425	0. 6425
12800	0. 8600	0. 8425	0. 8700	0. 8800	0. 8675	0. 8350	0. 8125	0. 7550

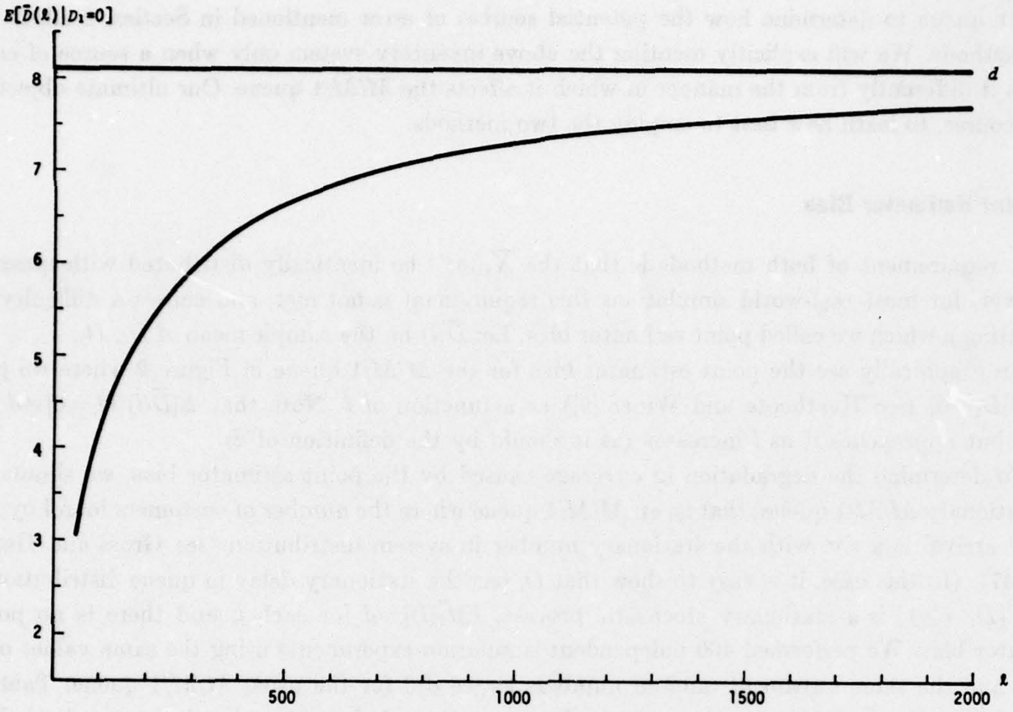


FIGURE 2. $E[\bar{\bar{D}}(l)|D_1=0]$ as a function of l for the $M/M/1$ queue with $\rho=0.9$.

B. Correlation Between the $\bar{X}_i(m)$'s

Suppose that $\{X_i, i\geq 1\}$ is a (covariance) stationary stochastic process. For $j=0, 1, 2, \ldots,$ let

$$C_j=\text{Cov}\,(X_i, X_{i+j})=c_{-j},$$

$$C_j(m)=\text{Cov}\,[\bar{X}_i(m), \bar{X}_{i+j}(m)]$$

($\bar{X}_i(m)$ is the sample mean of $X_{m(i-1)+1}, \ldots, X_{mi}$ for $i=1, 2, \ldots$), and

$$\rho_j(m)=C_j(m)/C_o(m).$$

Then it is easy to show that (see Mechanic and McKay [11])

$$C_j(m)=\sum_{i=-(m-1)}^{m-1} (1-|i|/m)\,C_{jm+i}/m.$$

A fundamental assumption of the method of batch means is that m is large enough so that the $\bar{X}_i(m)$*'s* are approximately uncorrelated, i.e., $\rho_j(m)\approx 0$ for $j\neq 0$. Let us examine the effect of violating this assumption. The true variance of $\bar{\bar{X}}(k, m)$ is given by (see Fishman [7])

$$\sigma^2[\bar{\bar{X}}(k, m)]=C_o(m)\left[1+2\sum_{j=1}^{k-1}(1-j/k)\rho_j(m)\right]\Big/k. \tag{4}$$

Denote the quantity in brackets by $a(k, m)$. It can be shown that (see, for example, Anderson [1], p. 448)

$$E\left\{\hat{\sigma}^2\left[\overline{\overline{X}}(k, m)\right]\right\}=[k-\alpha(k, m)]C_o(m)/k(k-1), \tag{5}$$

where $\hat{\sigma}^2[\overline{\overline{X}}(k, m)]$ was given by (2). Combining (4) and (5) we get

$$E\left\{\hat{\sigma}^2\left[\overline{\overline{X}}(k, m)\right]\right\}=b(k, m)\sigma^2\left[\overline{\overline{X}}(k, m)\right],$$

where $b(k, m)=\{[k/a(k, m)]-1\}/(k-1)$. Note that $b(k, m)=1$ when $\hat{\sigma}^2[\overline{\overline{X}}(k, m)]$ is an unbiased estimator.

Consider the sequence $\{D_i, i\geq 1\}$ for the stationary $M/M/1$ queue. We can compute C_j (which is positive) and, thus, $\rho_j(m)$ and $b(k,m)$ from formulas given by Daley [5]. Table 5 gives $\rho_j(m)(j=1, 2, 3)$ and $b(k, m)$ for the values of k and m previously considered. Observe that $\hat{\sigma}^2[\overline{\overline{X}}(k,m)]$ has a negative bias $(b(k,m)<1)$ when the $\rho_j(m)$'s are positive. To determine how large a degradation in coverage is caused by the bias, we repeated the simulation experiments of Subsection 4.A using the same random numbers. However, for the current experiments we divided each of the 400 variance estimates $\hat{\sigma}^2[\overline{\overline{X}}(k,m)]$ by the constant $b(k,m)$. Table 6 gives the resulting coverages. We conclude from a comparison of Tables 4 and 6 that the bias in $\hat{\sigma}^2[\overline{\overline{X}}(k,m)]$ can be a major source of error for batch means.

Consider now the sequence $\{E_i, i\geq 1\}$ for the stationary version of the above inventory system (i.e., X_1 has the stationary number in system distribution as given in [13], p. A48). We can compute C_j, which can be positive or negative, in a straightforward but laborious manner and we give $\rho_j(m)(j=1, 2, 3)$ and $b(k, m)$ in Table 7. Notice that the $\rho_j(m)$'s are now negative and that $\hat{\sigma}[\overline{\overline{X}}(k,m)]$ has a positive bias $(b(k, m)>1)$. This explains why the coverages (for batch means) in Table 2 are larger than 0.90. However, we believe that coverage greater than expected is not nearly as undesirable as coverage less than expected. Thus, correlation between the $\overline{X}_i(m)$'s would seem to be particularly troublesome only when it is positive.

TABLE 5. *$\rho_j(m)(j=1, 2, 3)$ and $b(k, m)$ for the Stationary $M/M/1$ Queue with $\rho=0.9$.*

m	$\rho_1(m)$	$\rho_2(m)$	$\rho_3(m)$	$b(5, m)$	$b(10, m)$	$b(20, m)$	$b(40, m)$
40	0.830	0.645	0.514				0.093
80	0.720	0.461	0.310			0.169	0.174
160	0.567	0.258	0.129		0.292	0.300	0.306
320	0.387	0.096	0.028	0.460	0.473	0.480	0.484
640	0.219	0.017	0.002	0.661	0.669	0.673	
1280	0.105	0.001	0.000	0.819	0.822		
2560	0.048	0.000	0.000	0.910			

TABLE 6. *Proportion of Coverage of Various 90% Confidence Intervals for d=8.1. in a Stationary M/M/1 Queue with ρ=0.9 and Variance Estimate* $\hat{\sigma}^2[\bar{\bar{X}}(k, m)]/b(k, m)$.

	Batch Means			
n \ k	5	10	20	40
1600	0.8175	0.8225	0.8450	0.8575
3200	0.8275	0.8500	0.8525	0.8600
6400	0.8400	0.8525	0.8725	0.8700
12800	0.8725	0.8575	0.8600	0.8850

TABLE 7. *($\rho_j(m)$)(j=1, 2, 3) and $b(k, m)$ for the Stationary (s, S) Inventory System with s=17, S=57, and Poisson Demand.*

m	$\rho_1(m)$	$\rho_2(m)$	$\rho_3(m)$	$b(5, m)$	$b(10, m)$	$b(20, m)$	$b(40, m)$
8	−0.188	−0.065	−0.029				2.444
16	−0.214	−0.037	−0.007			2.004	2.043
32	−0.188	−0.006	−0.000		1.595	1.615	1.626
64	−0.124	−0.000	−0.000	1.309	1.319	1.324	1.327
128	−0.071	−0.000	−0.000	1.160	1.162	1.164	
256	−0.038	−0.000	−0.000	1.081	1.082		
512	−0.020	−0.000	−0.000	1.041			

C. Nonnormality of the $\bar{X}_i(m)$'s

We can determine the effect of nonnormality on replication from Table 4, since the effect of the other potential source of error, point estimator bias, has been removed. Similarly, we can determine the effect of nonnormality on batch means from Table 6, since the effects of the other two sources of error, point estimator bias and the correlation between the $\bar{X}_i(m)$'s, have been removed. We conclude that for the models considered here, nonnormality is not a major source of error for either method, especially if k is approximately 20 or greater.

5. CONCLUSIONS AND FUTURE RESEARCH

In this paper we empirically showed that batch means is superior to replication for several well-known queueing and inventory systems. Although we cannot conclude from these results that batch means is superior for every stochastic system, we believe that it is a wise choice unless some additional information about the system being simulated is available. This and the other major conclusions of this paper are summarized below:

(a) The major source of error for replication is point estimator bias, i.e., $\mu(m) \neq \mu$.

(b) If m is chosen too small for replication, then coverage may decrease as k is increased.

(c) The major source of error for batch means is the bias in $\hat{\sigma}^2[\overline{\overline{X}}(k, m)]$ which is caused by the correlation between the $\overline{X}_i(m)$'s.

(d) Batch means appears to be superior to replication.

(e) If n is chosen too small, then the actual coverage of either method may be considerably lower than that desired, regardless of the choice of k.

(f) If it is possible to increase n for either method, then it is preferable (in terms of coverage) to hold k fixed and increase m (see Tables 1 and 2).

In order for replication to be a viable method for interval estimation, a procedure is needed for dealing with point estimator bias. To the best of our knowledge, no such generally applicable procedure exists at the present time. However, it should be mentioned that research related to this problem is currently being done by Ancker and Gafarian at the University of Southern California.

We have not explicitly discussed how to choose k and m for batch means. However, from Table 1 we know that for some systems and some values of n there will be no value of k which will produce coverage close to the desired level. We believe that what is needed is a sequential procedure which fixes k at a "reasonable" value and then successively increases m (and thus n) until the $\overline{X}_i(m)$'s are approximately uncorrelated. We will present such a procedure in the next paper in this series. Empirical results which will be reported there indicate that this sequential procedure works quite well for a variety of stochastic models. Thus, if properly implemented, batch means can be a viable method of interval estimation. For other research related to the choice of k and m, see Refs. [7] and [11].

ACKNOWLEDGMENTS

The author would like to thank John Carson and Frank Nick for their help in preparing this paper.

REFERENCES

[1] Anderson, T. W., *The Statistical Analysis of Time Series*, (Wiley, New York, 1970).

[2] Coveyou, R. R., and R. D. Macpherson, "Fourier Analysis of Uniform Random Number Generators," Journal of the Association for Computing Machinery *14*, 100–119 (1967).

[3] Crane, M. A., and D. L., Iglehart, "Simulating Stable Stochastic Systems, I; General Multi-Server Queues," Journal of the Association for Computing Machinery *21*, 103–113 (1974).

[4] Crane, M. A., and D. L. Iglehart, "Simulating Stable Stochastic Systems, II: Markov Chains," Journal of the Association for Computing Machinery *21*, 114–123 (1974).

[5] Daley, D. J., "The Serial Correlation Coefficients of Waiting Times in a Stationary Single Server Queue," Journal of the Australian Mathematical Society *8*, 683–699 (1968).

[6] Fishman, G. S., *Concepts and Methods in Discrete Event Simulation*, (Wiley, New York, 1973).

[7] Fishman, G. S., "Batch Means in Digital Simulation," Technical Report 75-7, curriculum in Operations Research and Systems Analysis, University of North Carolina at Chapel Hill (revised April 1976).

[8] Gross, D., and C. M. Harris, *Fundamentals of Queueing Theory*, (Wiley, New York, 1974).

[9] Heathcote, C. R., and P. Winer, "An Approximation for the Moments of Waiting Times," Operations Research *17*, 175–186 (1969).

[10] Iglehart, D. L., "Simulating Stable Stochastic Systems, V: Comparison of Ratio Estimators," Naval Research Logistics Quarterly *22*, 553–565 (1975).

[11] Mechanic, H., and W. McKay, "Confidence Intervals for Averages of Dependent Data in Simulations II," Report No. ASDD 17-202, IBM Corp., Yorktown Heights, New York, (1966).

[12] Ross, S. M., *Applied Probability Models with Optimization Applications*, (Holden-Day, San Francisco, 1970).

[13] Wagner, H. M., *Principles of Operations Research*, (Prentice-Hall, Englewood Cliffs, New Jersey, 1969).

COMPUTATION OF CONSTRAINED OPTIMUM QUANTITIES AND REORDER POINTS FOR TIME-WEIGHTED BACKORDER PENALTIES

John P. Matthews

University of Wisconsin—Madison
Madison, Wisconsin

ABSTRACT

The purpose of this paper and the accompanying tables is to facilitate the calculation of constrained optimum order quantities and reorder points for an inventory control system where the criterion of optimality is the minimization of expected inventory holding, ordering, and time-weighted backorder costs. The tables provided in the paper allow the identification of the optimal solution when order quantities and/or reorder points are restricted to a set of values which do not include the unconstrained optimal solution.

1. INTRODUCTION

There are many situations in which the analyst is forced to choose values of order points and order quantities from a finite set of alternatives. If stock can be ordered in multiples of say, m units, it would be under the most rare circumstances that an unconstrained optimal solution (UOS) would be equal to one of the admissible lot sizes. Similarly, it may not be possible to measure the inventory level at any other than, say, n different levels. For example, it may not be economical to measure accurately the exact level of content in pressurized gas containers. Thus, for practical purposes, the reorder point will be an integer number of containers. The tables in this paper allow the identification of the optimal solution when order point and order quantities are constrained.

The continuous review inventory system envisioned here is the same as the one examined by Holt, Modigliani, Muth, and Simon (HMMS) (Ref [4], p. 226). The assumptions concerning the inventory system follow.

- The lead time is shorter than the time between orders, the so-called "lot time."
- The order point or trigger level, T, is nonnegative.
- The lead time is constant and known.
- There is no serial correlation of sales rates between periods.
- An order not satisfied immediately from inventory is backlogged.
- The backorder penalty is a function of the time duration and amount of backorders.

The above assumptions allow the development of a total cost function which HMMS identify as Model Two (Ref. [3], Eq. 12–23). The cost function is equivalent to

$$K(Q,T)=C_F\frac{\bar{S}}{Q}+C_I\left(\frac{Q}{2}+T-\bar{S}_L\right)+\frac{\bar{S}_L}{Q}(C_I+C_D)\int_T^{\infty}\frac{(S_L-T)^2}{2S_L}f(S_L)\,dS_L, \tag{1}$$

where C_F is the ordering cost,

C_I is the holding cost/unit/year,

C_D is the backorder cost/unit/year,

$\bar{S}$ is the expected annual sales,

$\bar{S}_L$ is the expected sales over the lead time,

$f(S_L)$ is the probability density of S_L,

Q is the order quantity, and

T is the trigger level or reorder point.

Unfortunately, the values of Q and T which minimize $K(Q,T)$ are not easy to identify since, as *HMMS* express it, "the integral above is difficult to evaluate for many density functions of interest, . . ." in reference to the integral in (1).

When demand over lead time is normally distributed we may, with the aid of the accompanying table, evaluate the integral in (1) and employ a procedure presented in the paper to search for the optimal values (constrained or otherwise) of Q and T. Furthermore, the tables allow the $K(Q,T)$ cost surface to be easily generated.

2. COMMENTS ON RELATED PAPERS

The mathematical approach employed by Galliher, Morse, and Simond (GMS) [3] to obtain a total cost function is different from the one employed by HMMS. Their method of arriving at optimal values of Q and T is based upon an approximation to the cost function they derive. Deemer and Hoekstra [2] have developed tables which identify the optimal values of Q and T for the *GMS* model, but the tables do not facilitate cost evaluation or aid in the search for constrained optimal Q, T strategies. Koenigsberg [6] uses a model similar to that of GMS, but adopts a method which he states to be equivalent to minimizing holding and ordering costs subject to a fixed protection against shortage. Backorders are not time-weighted. Thatcher [8] uses a model in which stockholding costs are proportioned to the maximum amount of stock, which seems a doubtful approximation for many applications. Buckland [1] uses a nomogram to simplify the joint calculation of Q and I_t. Unfortunately, the construction of the nomogram is left to the reader. Backorders were not time-weighted in Buckland's treatment. Lampkin and Flowerdew [7] present an iterative procedure for the optimization of a related cost function, but require the generation of a table of values to be used in the optimization procedure.

Herron [4] generated a series of graphs suitable for identifying the UOS for the GMS model. The graphs in the Herron paper may be employed to assess the sensitivity of the minimum cost solution to changes in demand uncertainty over lead time. However, they cannot be used in the identification of a constrained optimal solution (COS), nor any sensitivity in cost to changes in the Q,T strategy from the UOS.

The paper proceeds in several sections. Section 3 displays the derivatives of $K(Q, T)$ and Section 4 describes the cost surface $K(Q, T)$. Section 5 describes properties of the isocost rings and the UOS. Section 6 discusses types of constraints that may be incurred and procedures to identify the COS. Section 7 illustrates the use of the tables and presents several examples. The tables and

optimization procedure presented here allow the constrained optimal solution and unconstrained optimal solution to be identified in several minutes of computations manually, or in seconds by computer. The tables allow the calculation of *Q* and *T* to within one-tenth of one standard deviation of demand over lead time.

3. PARTIAL DERIVATIVES OF K(Q, T)

Denote as D_{QQ} and D_{TT} the second partial derivatives of *K*(*Q*, *T*) with respect to *Q* and *T*, respectively. From (1), then, we have

$$D_{QQ}=\frac{2C_F\overline{S}}{Q^3}+\frac{2S_L}{Q^3}(C_I+C_D)\int_T^\infty\frac{(S_L-T)^2}{2S_L}f(S_L)\,dS_L, \tag{2}$$

$$D_{TT}=\frac{2\overline{S}_L}{Q}(C_I+C_D)\int_T^\infty\frac{1}{2S_L}f(S_L)\,dS_L. \tag{3}$$

Note that D_{QQ} and D_{TT} are greater than zero for all values of $\overline{S}_L$, C_F, C_I, C_D, and *Q* greater than zero.

4. THE TOTAL COST SURFACE

The isocost (*IC*) rings of the cost function *K*(*Q*, *T*) over the *Q*, *T* quadrant form nested rings whose size diminish as *K*(*Q*, *T*) decreases. The rings are symmetric about their major axis for reasons which will be discussed later in this paper. The major axis of the isocost ring is negative in slope, and the major axis of an *IC* ring is closer to the *Q*, *T* origin the lower the total cost value associated with the ring. Figure 1 shows an example where K_j represents total cost associated with ring *j*.

Since D_{QQ} and D_{TT} are both strictly positive, any point within an *IC* ring must yield a lower total cost value than any point *on* the *IC* ring thus ensuring that the *IC* rings are nested. As the reader may expect, the cost surface becomes relatively flat near the optimal solution. Since the *IC* rings are nested but not concentric, the rate of increase in *K*(*Q*, *T*) with respect to movement

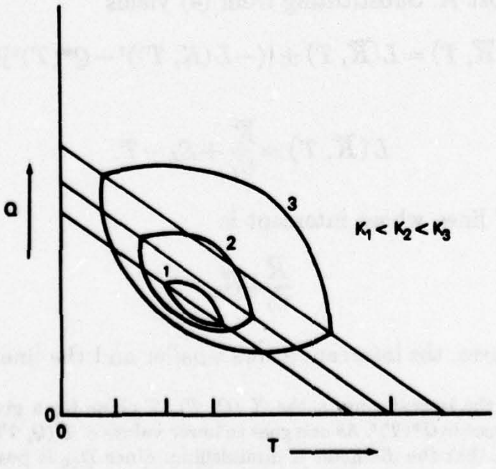


FIGURE 1. IC ring.

from the *UOS* is very sensitive to the direction of movement. Thus, when it is not possible to implement the *UOS*, great care must be exercised in choosing among alternative constrained solutions.

5. PROPERTIES OF THE UNCONSTRAINED OPTIMAL SOLUTION

It will be shown that the *UOS*, under conditions discussed later in the paper, will lie at the point of tangency of one line of a family of parallel lines and a curve convex to the Q, T origin.

Denote as $Q^*(T)$ that value of Q which, for a given value of T, minimizes $K(Q, T)$. An expression for $Q^*(T)$ may be found by setting the first derivative of (1) with respect to Q equal to zero and solving for Q. Doing so yields

$$Q^*(T)=[C_1+C_2EBP(T)]^{1/2}, \tag{4}$$

where $C_1=2C_P\bar{S}/C_I$, $C_2=2\bar{S}_L(C_I+C_D)/C_I$, and $EBP(T)$ represents the integral in (1). When T is large, $EBP(T)$ is small and $Q^*(T)$ approaches the familiar Wilson lot-size formula. From (1), it follows easily that $EBP(T)$ and therefore $Q^*(T$ are monotonically decreasing functions of T.

Denote the *UOS* to (1) as Q^{**}, T^{**} and the optimal cost as K^*. Since $Q^*(T)$ is a single-valued function of T, it follows that Q^{**} must lie on the curve defined in (4), i.e., $Q^*(T^{**})=Q^{**}$. It will be assumed in the discussion to follow that $Q^*(T)$ is convex, although the assumption is not supported by a proof. $Q^*(T)$ has always been found to be convex for all values assigned to C_1 and C_2 by the author. To ensure that $Q^*(T)$ is convex for a particular problem, the curve may be traced out for selected values of T and $EBP(T)$ with the aid of the tables. Under the assumption of convexity of $Q^*(T)$, the optimal solution will be shown to be unique.

Consider the isocost rings over the Q, T plane. Setting $K(Q, T)$ equal to some constant, say $\bar{K}$, and solving (1) for Q, we obtain

$$Q(\bar{K}, T)=\frac{\bar{K}}{C_I}+\bar{S}_L-T\pm\left[\left(T-\bar{S}_L-\frac{\bar{K}}{C_I}\right)^2-C_1-C_2EBP(T)\right]^{1/2} \tag{5}$$

Thus, for a given value of $\bar{K}$ and T, (5) yields the value(s) of Q for a given value of T on the *IC* ring associated with total cost $\bar{K}$. Substituting from (4) yields

$$Q(\bar{K}, T)=L(\bar{K}, T)\pm[(-L(K, T))^2-Q^*(T)^2]^{1/2}, \tag{6}$$

where

$$L(\bar{K}, T)=\frac{\bar{K}}{C_I}+S_L-T. \tag{7}$$

$L(\bar{K}, T)$ is a family of lines whose intercept is

$$\frac{\bar{K}}{C_I}+\bar{S}_L$$

and slope is -1. As $\bar{K}$ decreases, the intercept grows smaller and the lines move toward the origin.*

*Note that the diameter of the isocost ring in the K (Q, T), T plane for a given value of T is related to the amount by which $(-L(K, T))^2$ exceeds $Q^*(T)^2$. As one goes to lower values of $K(Q, T)$ holding T constant, $-L(K, T)$ approaches $Q^*(T)$, thus ensuring that the diameter is diminishing. Since D_{QQ} is positive for all Q and T, it follows that the *IC* rings are nested, since all points within (outside) the ring yield cost values less (more) than all points on the ring.

In the upper frame of Figure 2, three members of the family of lines $Q=L(K, T)$ are drawn, with $L(K^*, T)$ denoting the line associated with the minimum value of $K(Q, T)$. For values of T such that $L(\overline{K}, T)$ exceeds $Q^*(T)$, it follows from (6), as shown in Frame a of Figure 2, that the isocost ring has two values of $Q(\overline{K}, T)$. Thus it follows from (6) that for values of $\overline{K}$ and T such that

(8) $L(\overline{K}, T)>Q^*(T)$, $Q(\overline{K}, T)$ has two real solutions,

(9) $L(\overline{K}, T)=Q^*(T)$, $Q(\overline{K}, T)$ has one real solution,

(10) $L(\overline{K}, T)<Q^*(T)$, $Q(\overline{K}, T)$ has no real solution.

If, for a given T, say T_k, $Q(\overline{K}, T_k)$ has two real solutions, say Q_1 and Q_2, then by definition of the IC ring,

(11) $K(Q_1, T_k)=K(Q_2, T_k)=\overline{K}$.

But since $D_{QQ}>0$, it follows that there must be a λ, $0<\lambda<1$, such that

(12) $K(Q_\lambda, T_k)<\overline{K}$

where $Q_\lambda=\lambda Q_1+(1-\lambda)Q_2$. Thus, $\overline{K}$ cannot be optimum if for some T it is found that $L(\overline{K}, T)>Q^*(T)$. The value of $\overline{K}$ which corresponds to the line tangent to $Q^*(T)$ yields one value of $Q(\overline{K},T)$ for one value of T and imaginary values of $Q(\overline{K}, T)$ for all other values of T. Since all lines above the tangent line lie above $Q^*(T)$ for some value of T and therefore yield two values of $Q(\overline{K}, T)$, those lines must be associated with a value of $\overline{K}$ greater than K^*. All lines below the line of tangency do not intersect $Q^*(T)$ and therefore do not yield real valued solutions. For $Q^*(T)$ convex, one and only one line in the family of $L(\overline{K}, T)$ lines will be tangent to $Q^*(T)$. Therefore, the point of tangency between that line and the $Q^*(T)$ must be the optimal solution.

With the aid of the tables, the convexity of $Q^*(T)$ for a given problem may be investigated and the UOS easily identified. An iterative procedure which will identify the unconstrained optimal solution in seven iterations is provided in the appendix.

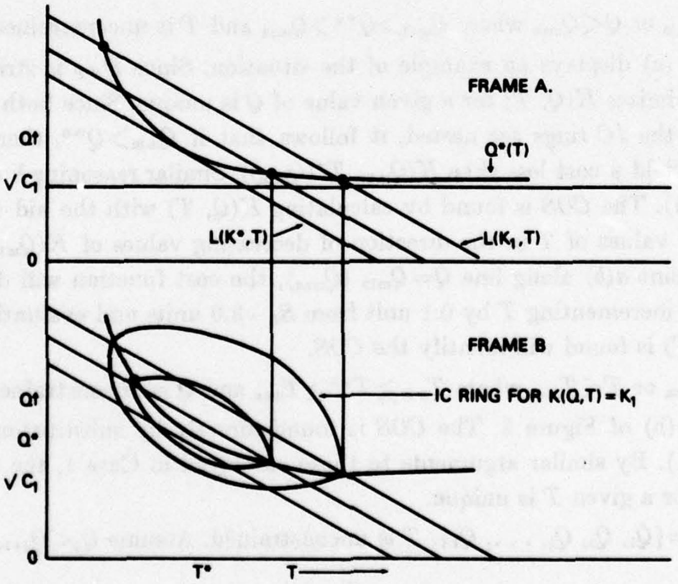


FIGURE 2

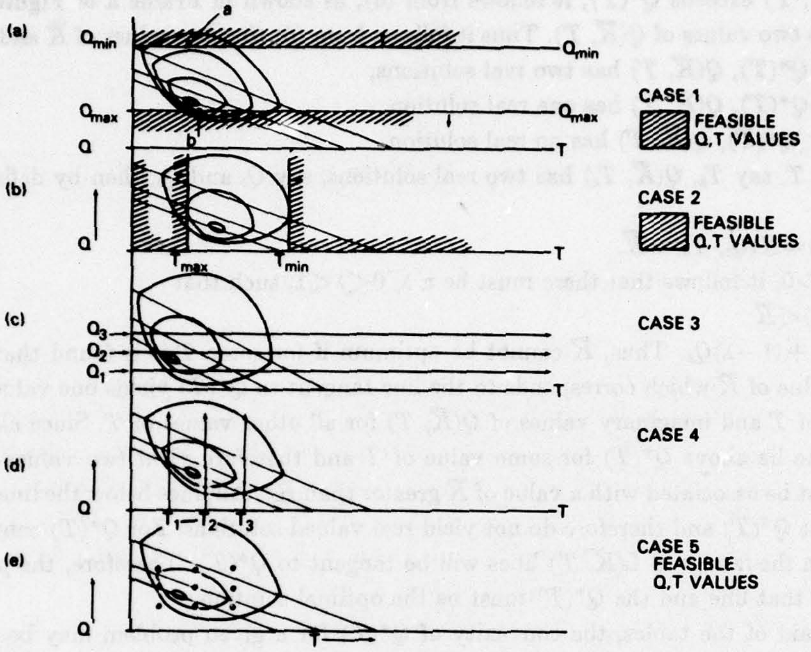


FIGURE 3.

6. FORMS OF CONSTRAINTS ON Q AND T

CASE 1: $Q > Q_{min}$ or $Q < Q_{max}$ where $Q_{min} > Q^{**} > Q_{max}$ and T is unconstrained.

Figure 3 Frame (a) displays an example of the situation. Since D_{TT} is strictly positive, the value of T which minimizes $K(Q, T)$ for a given value of Q is unique. Since both D_{QQ} and D_{TT} are strictly positive and the *IC* rings are nested, it follows that if $Q_{min} > Q^{**}$, there is no Q greater than Q_{min} which will yield a cost less than $K(Q_{min}, T^*(Q_{min}))$. Similar reasoning holds for $Q^{**} > Q_{max}$ and $K(Q_{max}, T^*(Q_{max}))$. The *COS* is found by calculating $K(Q, T)$ with the aid of the tables and evaluating successive values of T in the direction of decreasing values of $K(Q_{min}, T)$. Clearly, as one moves toward point $a(b)$ along line $Q = Q_{min}$ (Q_{max}), the cost function will decrease until the *COS* is passed. Thus, incrementing T by 0.1 unit from $S_L - 3.0$ units and evaluating $K(Q, T)$ until an increase in $K(Q, T)$ is found will identify the *COS*.

CASE 2: $T > T_{min}$ or $T < T_{max}$ where $T_{min} \geq T^{**} \geq T_{max}$ and Q is unconstrained.

Refer to Frame (b) of Figure 3. The *COS* is found directly by substituting the value of T min (or T max) in (4). By similar arguments to those employed in Case 1, the value of Q which minimized $K(Q, T)$ for a given T is unique.

CASE 3: $Q = \Phi = \{Q_1, Q_2, Q_3, \ldots, Q_k\}$, T is unconstrained. Assume $Q_j < Q_{j+1}$, $j = 1, \ldots, k-1$.

Refer to Frame (c) of Figure 3. It is apparent from Figure 3 that as successive values of $K(Q_j, T^*(Q_j))$ are evaluated, the first nondecreasing value of $K(Q_j, T^*(Q_j))$ indicates that the optimal constrained solution has been passed. Starting with $j=1$, calculate $K(Q_j, T^*(Q_j))$ as in Case 1. Continue to increment j until for some w,

$$K(Q_{w-1}, T^*(Q_{w-1}))\geq K(Q_w, T^*(Q_w))\leq K(Q_{w+1}, T^*(Q_{w+1})).$$

Since the isocost rings are nested and D_{QQ} and D_{TT} are strictly positive, it follows that Q_w, $T^*(Q_w)$ is the *COS*.

CASE 4: $T\epsilon\tau=\{T_1, T_2, T_3, \ldots, T_l\}$, Q is unconstrained.

Refer to Frame (d) of Figure 3. Assume $T_j<T_{j+1}, j=1, \ldots, l-1$. It is apparent from Figure 3 that as successive values of $K(Q^*(T_j), T_j)$ are evaluated, the first nondecreasing value of $K(Q^*(T_j), T_j)$ indicates that the optimal solution has been passed. Starting with $j=1$ calculate $K(Q^*(T_j), T_j)$ as in Case 2. Continue to increment j and calculate $K(Q^*(T_j), T_j)$ until for some V,

$$K(Q^*(T_{V-1}), T_{V-1})\geq K(Q^*(T_V), T_V)\leq K(Q^*)T_{V+1}), T_{V+1}).$$

If we employ arguments similar to Case 3, it follows that $Q^*(T_V)$, T_V is the *COS*.

CASE 5: $Q\epsilon\phi$, $T\epsilon\tau$.

Refer to Frame (c) of Figure 3. If the number of feasible Q, T strategies is small, each of the $k\times l$ points may be evaluated. If the number is large, the cost surface may be generated to visually locate the *COS*. Figure 4 indicates the flow of the computations which will generate the $K(Q, T)$ surface from $Q_{\min}$ to $Q_{\max}$ and for T from $\overline{S}_L-3$ to $\overline{S}_L+3$.

7. USE OF THE TABLES

In order to be applicable to a specific problem, the units of measure must be standardized; therefore, all measurements are in terms of standard deviations. Thus, an annual sales rate of 1000 units, a lead time of 0.08 year, an order point of 90, and a standard deviation of 10 units would yield parameters as follows:

$$\overline{S}=100,\ \overline{S}_L=8, \text{ and } T=9$$

The C_F, C_I, and C_D would be in dimensions of \$/order, \$/σ·year, and \$/σ·year, respectively. Thus, for a T of 9 and $\overline{S}_L$ of 8, the corresponding time-weighted value is $EBP(T)=0.003700$. The interpretation is that for every order cycle we expect on the average to accumulate

$$0.\left(003700 * \frac{10 \text{ units}}{\sigma} * \frac{29 \text{ days}}{\text{lead time}}\right)$$

or 1.07 unit days of backorders.

Continuing, let us assume that the cost parameters of the problem are as follows: $C_I=\$100/\sigma$ year, $C_F=\$200$, $C_D=\$40{,}000/\sigma$ year. We obtain $C_1=400$ and $C_2=6416$. If the procedure presented in the appendix were employed, the UOS would be found to be

$$Q^{**}=20.465 \text{ and } T^{**}=9.1. \tag{13}$$

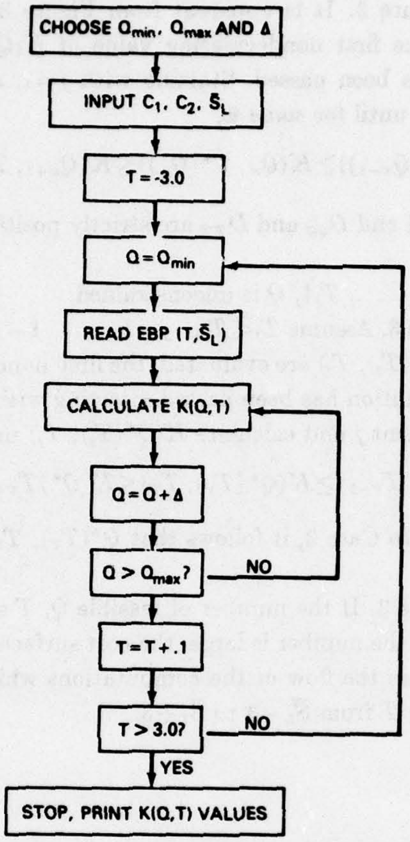

CHOOSE Q_{min}, Q_{max} AND Δ
INPUT C_1, C_2, S_L
T = -3.0
Q = Q_{min}
READ EBP (T, $\bar{S}_L$)
CALCULATE K(Q,T)
Q = Q + Δ
Q > Q_{max}?
NO
T = T + .1
T > 3.0?
NO
YES
STOP, PRINT K(Q,T) VALUES

FIGURE 4. Flow chart to map the K (Q, T) surface.*

*The evaluation of 441 points on the $K(Q, T)$ surface has been found to require less than 1 second of CPU time on a Univac 1110.

Other, more complex constraining relationships would best be examined by observing the location of feasible Q, T values on the $K(Q, T)$ surface.

Table 1 displays the results of applying the techniques discussed in Section 6 in solving the problem above and constraining the solution by several example methods. The values in the Q and T columns are cost-minimizing values unless otherwise constrained. For example, for Case 1, a trigger level of 8.8 will minimize $K(Q,T)$, given Q must not be less that 40 and T is unconstrained.

Note that if the values of Q and T were rounded down from the *UOS* to (for Case 5) the next smallest feasible values of Q=18 and T=8, the resulting expected cost of that strategy would be 12% higher than the *COS* of Q=22 and T=10. Rounding Q up to 22 and T down to 8 is slightly better, costing 8% more than the *COS*. It is clear that simply rounding the *UOS* values up or down is not necessarily going to yield a very satisfactory solution to the constrained problem.

TABLE 1.

Case	Constraint	Values		COS?	Cost
		Q	T		
1.	$Q \geq 40$, T unconstrained	40.0	8.8	yes	2626.30
2.	$T \geq 10.5$, Q unconstrained	20.0	10.5	yes	2250.00
3.	$Q \epsilon \phi = \{18, 22, 26\}$	18.0	9.2	no	2172.33
	T unconstrained	22.0	9.1	yes	2162.90
		26.0	9.0	no	2214.88
4.	$T \epsilon \tau = \{6, 8, 10\}$	47.3	6.0	no	4530.00
	Q unconstrained	23.8	8.0	no	2380.00
		20.0	10.0	yes	2200.00
5.	$Q \epsilon \phi$, $T \epsilon \tau$	18.0	6.0	no	6914.91
		18.0	8.0	no	2477.67
		18.0	10.0	no	2215.81
		22.0	6.0	no	5984.92
		22.0	8.0	no	2390.83
		22.0	10.0	yes	2212.94
		26.0	6.0	no	5402.63
		26.0	8.0	no	2392.24
		26.0	10.0	no	2272.49

BIBLIOGRAPHY

[1] Buckland, J. C. L., "A Nomogram for Stock Control," Operational Research Quarterly *20* (4) 445 (1969).

[2] Deemer, R. L., and D. Hoekstra, "Improvement of M.I.T. Non-Reparables Model," United States Army Logistics Management Center Publication, Fort Lee, Virginia (1968).

[3] Galliher, H. P., P. M. Morse, and M. Simond, "Dynamics of Two Classes of Continuous-Review Inventory Systems", Operations Research *7* (3) (1959).

[4] Herron, D. P., "Use of Dimensionless Ratios to Determine Minimum-Cost Inventory Quantities," Naval Research Logistics Quarterly 167–175 (July 1966).

[5] Holt, Charles C., Franco Modigliani, John F. Muth, and Herbert A. Simon, *Planning Production, Inventories, and Work Force.* (Prentice-Hall, Inc.: Englewood Cliffs, N.J., 1960).

[6] Koenigsberg, E., "On a Multiple Re-order Point Inventory Policy," Operational Research Quarterly *12*, 27 (1961).

[7] Lampkin, W., and A. D. J. Flowerdew, "Computation of Optimum Re-order Levels and Quantities for a Re-order Level Stock Control System," Operational Research Quarterly *14*, (3) 263 (1963).

[8] Thatcher, A. R., "Some Results on Inventory Problems," Journal of the Royal Statistical Society B *24*, (1) (1962).

APPENDIX

Search Procedure for T^{**}

It follows from (10) that as $\overline{K}$ increases, $L(\overline{K}, T)$ moves outward from the origin. For $Q^*(T)$ convex, the points of intersection of a line $L(\overline{K}, T)$ and the curve $Q^*(T)$ move apart and away from T^{**} as $\overline{K}$ increases. Thus, if an increase in T from, say, T_m to T_l along the curve $Q^*(T)$ results in an increase such that $K(Q^*(T_l),T_l) > K(Q^*(T_m),T_m)$, then all points beyond T_l may be eliminated from consideration in the search for T^{**}. Conversely, if a decrease in T from, say, T_q to T_p results in an increase such that $K(Q^*(T_l),T_p) > K(Q^*(T_q),T_q)$, then all points beyond T_q may be eliminated from consideration. The search procedure makes use of these observations.

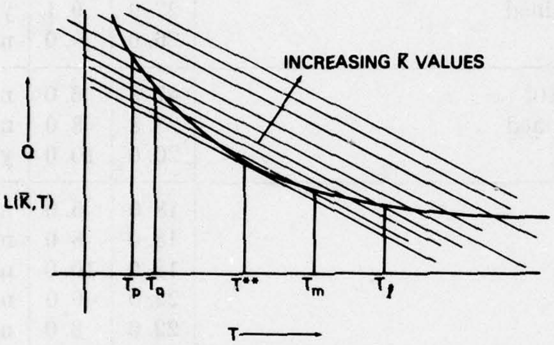


FIGURE A1.

For each value of $\overline{S}_L$, there are 61 tabled values of $T - \overline{S}_L$. The search procedure will be used to identify the tabled value of T and $Q^*(T)$ which minimizes (1). Let n, $1 \leq n \leq 61$ denote a row of the table. Let $T_j = \overline{S}_L - 3.1 + (0.1)\ N_j$ where N_j denotes a value of n. Let $K_j = K(Q^*(T_j),\ T_j)$ where $Q^*(T_j)$ is determined in (4).

The search procedure at each iteration eliminates from further consideration sets of values of n. At the start of each iteration, the uneliminated values of n are divided into three mutually exclusive sets. Given N_1 and N_2, sets S_1, S_2, and S_3 are formed.

If $N_1 < N_2$, S_1 contains values of $n \leq N_1$,
S_2 contains values of $n \geq N_2$,
S_3 contains values of $n > N_1$ and $< N_2$.

If $N_1 > N_2$, S_1 contains values of $n \geq N_1$,
S_2 contains values of $n \leq N_2$,
S_3 contains values of $n > N_2$ and $< N_1$.

The procedures may now be presented.

STEP 1. Set $N_1 = 24$, $N_2 = 38$.

STEP 2. Form sets S_1, S_2, and S_3 as indicated above.

STEP 3. Calculate K_1 and K_2.

STEP 4. If $K_1 > K_2$, eliminate set S_1 from further consideration and set N_1 equal to N_2. If $K_1 < K_2$, eliminate set S_2 from further consideration.

Let the larger of the two remaining sets be denoted as S_0.

ITERATION 1. If $\forall\, n\epsilon S_0$, $n>N_1$, then $N_2=N_1+10$, otherwise $N_2=N_1-10$. Perform Steps 2 through 4 and proceed to Iteration 2.

ITERATION 2. If $\forall\, n\epsilon S_0$, $n>N_1$, then $N_2=N_1+4$, otherwise $N_2=N_1-4$. Perform Steps 2 through 4 and proceed to Iteration 3.

ITERATION 3. If $\forall\, n\epsilon S_0$, $n>N_1$, then $N_2=N_1+6$, otherwise $N_2=N_1-6$. Perform Steps 2 through 4 and proceed to Iteration 4.

ITERATION 4. If $\forall\, n\epsilon S_0$, $n>N_1$, then $N_2=N_1+2$, otherwise $N_2=N_1-2$. Perform Steps 2 through 4 and proceed to Iteration 5.

ITERATION 5. If $\forall\, n\epsilon S_0$, $n>N_1$, then $N_2=N_1+2$, otherwise $N_2=N_1-2$. Perform Steps 2 through 4 and proceed to Iteration 6.

ITERATION 6. There are 3 values that remain, one on each side of the present value of N_1. Set N_2 equal to N_1+1. Let S_2 consist only of the N_2. S_3 is null. S_1 consists of the remaining two values. Perform Steps 3 and 4 and proceed to Iteration 7.

ITERATION 7. If S_0 contains one value of n, go to Step 5. If S_0 contains 2 values, let $N_2=N_1-1$. Perform Steps 3 and 4 and proceed to step 5.

STEP 5. Only one value remains. All other values of n yield higher costs, thus the present value of N_1 is the optimal value.

Table A1 displays the rate at which values are eliminated as the search progresses.

Table A2 presents the results of application of the search procedure to the example problem presented in Section 7.

TABLE A1

j	0	1	2	3	4	5	6	7
Values eliminated at iteration j	24	14	10	4	4	2	1 or 2	1 or 0
Total values eliminated	24	38	48	52	56	58	59 or 60	60

TABLE A2

Iteration	N_1	N_2	S_1	S_3	S_2	K_1	K_2	Eliminate
0	24	38	1–24	25–37	38–61	2789.90	2181.01	S_1
1	38	48	25–38	39–47	48–61	2181.01	2180.08	S_1
2	48	52	39–48	49–51	52–61	2180.08	2241.04	S_2
3	48	42	48–51	43–47	39–42	2180.08	2156.55	S_1
4	42	44	39–42	43	44–47	2156.55	2158.88	S_2
5	42	40	42, 43	41	39, 40	2156.55	2162.96	S_2
6	42	43	41, 42	-------	43	2156.55	2156.76	S_2
7	42	41	42	-------	41	2156.55	2158.49	S_2

Thus, the *UOS* is $Q^{**}=20.465$ and $T^{**}=9.1$.

$T-\bar{S}_L$	$\bar{S}_L$																								
	2.0	2.5	3.0	3.5	4.0	5.0	6.0	7.0	8.0	9.0	10.0	12.0	14.0	16.0	18.0	20.0	25.0	30.0	35.0	40.0	50.0	60.0	70.0	85.0	100.
−3.0	2.320050	1.812819	1.500188	1.290055	1.135037	.918395	.772826	.667747	.588134	.525647	.475257	.398931	.343812	.302118	.269464	.243193	.195558	.163541	.140539	.123212	.098844	.082525	.070832	.058417	.049705
−2.9	2.159283	1.690287	1.402333	1.207582	1.063348	.861257	.725192	.626853	.552285	.493724	.446477	.374877	.323146	.283999	.253332	.228655	.183897	.153807	.132184	.115894	.092981	.077634	.066636	.054959	.046764
−2.8	2.004933	1.572805	1.308426	1.128261	.994337	.806201	.679267	.587413	.517703	.462923	.418705	.351660	.303195	.266506	.237757	.214617	.172637	.144405	.124114	.105826	.087317	.072909	.062584	.051618	.043923
−2.7	1.857008	1.460393	1.218415	1.052083	.927999	.753224	.635052	.549427	.484387	.433242	.391940	.329280	.283960	.249639	.222737	.201079	.161776	.135336	.116330	.102006	.081853	.068351	.058673	.048395	.041182
−2.6	1.715520	1.352970	1.132264	.979035	.864326	.702321	.592542	.512892	.452334	.404682	.366180	.307735	.265440	.233396	.208272	.188040	.151314	.126600	.108830	.095437	.076589	.063959	.054905	.045290	.038540
−2.5	1.580489	1.251000	1.049939	.909102	.803309	.653488	.551734	.477804	.421542	.377239	.341423	.287025	.247634	.217778	.194361	.175500	.141250	.118195	.101615	.089115	.071523	.059732	.051280	.042301	.035998
−2.4	1.451941	1.154319	.971404	.842270	.744938	.606717	.512624	.444162	.392009	.360911	.317668	.267146	.230540	.202782	.181003	.163457	.131585	.110123	.094683	.083043	.066656	.055672	.047796	.039429	.033556
−2.3	1.329924	1.062761	.896623	.778519	.689200	.562001	.475206	.411959	.363730	.235695	.294912	.248098	.214157	.188408	.168198	.151911	.122316	.102380	.088036	.077218	.061987	.051776	.044454	.036674	.031212
−2.2	1.214511	.976136	.825559	.717826	.636079	.519332	.439473	.381192	.336702	.301588	.273152	.229878	.198482	.174652	.155942	.140859	.113443	.094968	.081670	.071641	.057517	.048045	.041253	.034035	.028968
−2.1	1.105469	.394314	.758168	.660166	.585557	.478696	.405147	.351853	.310918	.278585	.252383	.212482	.183513	.161514	.144235	.130301	.104965	.870884	.075587	.066310	.053243	.044479	.038193	.031513	.026822
−2.0	1.004244	.817181	.694403	.605511	.537613	.440080	.373027	.323934	.286373	.256679	.232600	.195907	.169247	.148990	.133073	.120234	.096879	.801127	.069784	.061225	.049166	.041076	.035273	.029105	.024774
−1.9	.910358	.744629	.634213	.553826	.492222	.403468	.342291	.297425	.263057	.235864	.213798	.180147	.155679	.137077	.122455	.110656	.089184	.074696	.064259	.056383	.045384	.037836	.032493	.026813	.022824
−1.8	.823325	.676547	.577542	.505072	.449354	.368840	.313194	.272314	.240962	.216132	.195969	.165197	.142804	.125770	.112375	.101562	.081876	.068588	.059012	.051784	.041596	.034758	.029851	.024635	.020970
−1.7	.742603	.612826	.524327	.459207	.408977	.336173	.285719	.248587	.220074	.197471	.179104	.151049	.130616	.115065	.102829	.092949	.074954	.062801	.054040	.047426	.038100	.031840	.027347	.022570	.019214
−1.6	.667834	.553350	.474498	.416180	.371051	.305440	.259845	.226228	.200381	.179872	.163193	.137695	.119109	.104955	.093813	.084813	.068413	.057331	.049341	.043306	.034796	.029081	.024979	.020617	.017552
−1.5	.598724	.497998	.427980	.375937	.335533	.276612	.235549	.205218	.181866	.163318	.148222	.125125	.108274	.095433	.085319	.077147	.062248	.052175	.044910	.039422	.031679	.026480	.022746	.018776	.015985
−1.4	.535005	.446645	.384690	.338416	.302375	.249652	.212804	.185534	.164510	.147795	.134179	.113328	.098102	.086491	.077342	.069946	.056455	.047330	.040745	.035770	.038749	.024033	.020646	.017644	.014512
−1.3	.476422	.399157	.344538	.303548	.271521	.224522	.191579	.167152	.148293	.133283	.121047	.102290	.088581	.078119	.069871	.063201	.051028	.042789	.036842	.032347	.026003	.021739	.018677	.015420	.013130
−1.2	.422728	.355396	.307427	.271260	.242910	.201178	.171840	.150043	.133190	.119762	.108806	.091996	.079698	.070306	.062898	.056904	.045959	.038548	.033196	.029149	.023436	.019596	.016837	.013902	.011838
−1.1	.373680	.315214	.273253	.241470	.216478	.179571	.153548	.134175	.119174	.107209	.097437	.082430	.071439	.063040	.056411	.051045	.041242	.034599	.029801	.026171	.021046	.017599	.015123	.012487	.010634
−1.0	.329034	.278459	.241904	.214090	.192150	.159648	.136661	.119513	.106216	.095596	.086916	.073572	.063789	.056307	.050399	.045614	.036867	.030937	.026651	.023408	.018827	.015746	.013532	.010174	.009517
−.9	.288550	.244974	.213262	.189027	.169850	.141350	.121133	.106019	.904281	.084896	.077218	.065402	.056730	.050092	.044847	.040598	.032825	.027552	.023739	.020854	.016776	.014032	.012060	.009960	.008483
−.8	.251985	.214594	.187204	.166181	.140494	.124615	.106912	.093650	.083334	.075076	.068313	.057895	.050241	.044378	.039742	.035984	.029106	.024437	.021059	.018502	.014887	.012454	.010705	.008842	.007531
−.7	.219097	.187151	.163600	.145447	.130993	.109376	.093964	.082361	.073336	.066102	.060173	.051028	.044302	.039146	.035066	.031757	.025697	.021581	.018601	.016345	.013154	.011006	.009461	.007815	.006657
−.6	.189645	.162472	.142317	.126716	.114257	.095562	.082176	.072104	.064245	.067938	.052764	.044774	.038891	.034376	.030802	.027902	.022586	.018974	.016358	.014376	.011572	.009684	.008325	.006878	.005859
−.5	.163391	.140384	.123217	.109875	.099187	.083099	.071543	.062829	.056018	.050546	.046052	.039104	.033982	.030049	.026932	.024402	.019761	.016606	.014319	.012586	.010134	.008481	.007292	.006025	.005133
−.4	.140099	.120711	.106161	.094807	.085685	.071909	.061983	.054482	.048609	.043885	.040000	.033989	.029551	.026140	.023436	.021240	.017209	.014465	.012475	.010967	.008832	.007393	.006357	.005253	.004476
−.3	.119535	.103276	.091007	.081395	.073649	.061915	.053433	.047008	.041970	.037912	.034571	.029396	.025571	.022623	.020294	.018396	.014912	.012538	.010816	.009510	.007660	.006413	.005515	.004558	.003884
−.2	.101472	.087907	.077614	.069510	.062977	.053034	.045824	.040351	.036052	.032584	.029727	.025295	.022015	.019489	.017434	.015853	.012856	.010813	.009330	.008205	.006611	.005535	.004761	.003935	.003353
−.1	.085690	.074430	.065842	.059061	.053565	.045186	.039091	.034454	.030805	.027858	.025427	.021651	.018854	.016697	.014984	.013590	.011022	.009276	.008006	.007042	.005675	.004753	.004088	.003380	.002881
−.0	.071975	.062678	.055551	.049903	.045312	.038290	.033166	.029259	.026179	.023688	.021631	.018432	.016059	.014228	.012773	.011587	.009406	.007916	.006834	.006012	.004846	.004059	.003492	.002387	.002461
.1	.060124	.052490	.046608	.041930	.038116	.032265	.027982	.024709	.022124	.020030	.018300	.015605	.013604	.012058	.010828	.009826	.007980	.006718	.005801	.005104	.004116	.003448	.002967	.002453	.002091
.2	.049942	.043708	.038882	.035029	.031879	.027033	.023474	.020747	.018590	.016841	.015393	.013137	.011458	.010160	.009127	.008285	.006732	.005669	.004896	.004309	.003475	.002912	.002506	.002072	.001767
.3	.041246	.036184	.032247	.029093	.026507	.022516	.019576	.017318	.015529	.014077	.012873	.010994	.009595	.008512	.007649	.006945	.005646	.004756	.004109	.003617	.002918	.002445	.002105	.001741	.001484
.4	.033864	.029777	.026585	.024018	.021903	.018542	.016228	.014370	.012895	.011698	.010701	.009146	.007987	.007088	.006372	.005787	.004707	.003967	.003428	.003018	.002435	.002041	.001757	.001454	.001239
.5	.027635	.024356	.021783	.019707	.017996	.015339	.013370	.011850	.010642	.009658	.008841	.007563	.006608	.005867	.005276	.004793	.003901	003288	.002842	.002503	.002020	.001694	.001458	.001206	.001029
.6	.022415	.019799	.017738	.016070	.014691	.012543	.010946	.009712	.008728	.007926	.007259	.006214	.005433	.004826	.004371	.003945	.003212	002789	.002842	.002063	.001666	.001397	.001203	.000995	.000849
.7	.018066	.015993	.014352	.013020	.011916	.010191	.008905	.007908	.007113	.006463	.005922	.005074	.004438	.003944	.003556	.003227	.002629	002218	.001918	.001690	.001365	.001145	.000986	.000816	.000696
.8	.014469	.012836	.011538	.010481	.009603	.008227	.007197	.006398	.005759	.005236	.004800	.004116	.003602	.003203	.002883	.002622	.002137	001804	.001560	.001375	.001111	.000932	.000803	.000664	.000567
.9	.011512	.010235	.009215	.008381	.007687	.006597	.005779	.005142	.004632	.004214	.003865	.003317	.002905	.002584	.062327	.002116	.001726	001458	.001261	.001112	.000898	.000754	.000649	.000537	.000459
1.0	.009100	.008106	.007310	.006658	.006113	.005255	.004608	.004104	.003700	.003368	.003091	.002655	.002326	.002070	.001865	.001697	.001385	.001170	.001013	.000893	.000722	.000606	.000522	.000432	.000369
1.1	.007144	.006377	.005760	.005252	.004828	.004157	.003650	.003254	.002936	.002674	.002455	.002110	.001851	.001648	.001485	.001351	.001104	.000932	.000807	.000714	.000576	.000483	.000416	.000345	.000294
1.2	.005571	.004982	.004507	.004115	.003786	.003265	.002871	.002562	.002313	.002108	.001937	.001666	.001462	.001302	.001174	.001069	.000873	.000738	.000639	.000564	.000456	.000383	.000330	.000273	.000233
1.3	.004314	.003865	.003502	.003201	.002949	.002547	.002242	.002003	.001809	.001650	.001517	.001306	.001147	.001022	.000922	.000839	.000686	.000580	.000503	.000443	.000359	.000301	.000260	.000215	.000184
1.4	.003317	.002978	.002702	.002473	.002280	.001973	.001739	.001554	.001405	.001283	.001180	.001016	.000893	.000796	.000718	.000654	.000535	.000453	.000392	.000346	.000280	.000235	.000203	.000168	.000144
1.5	.002533	.002278	.002070	.001897	.001750	.001517	.001338	.001198	.001084	.000996	.000911	.000785	.000690	.000616	.000556	.000507	.000415	.000351	.000304	.000268	.000217	.000183	.000157	.000130	.000111
1.6	.001920	.001729	.001574	.001444	.001334	.001158	.001023	.000916	.000830	.000758	.000698	.000602	.000530	.000473	.000427	.000389	.000319	.000270	.000234	.000207	.000167	.000141	.000121	.000100	.000086
1.7	.001445	.001304	.001188	.001091	.001009	.000877	.000776	.000695	.000630	.000576	.000531	.000458	.000403	.000360	.000325	.000297	.000243	.000206	.000179	.000158	.000128	.000107	.000093	.000077	.000066
1.8	.001079	.000975	.000890	.000818	.000757	.000659	.000584	.000524	.000475	.000436	.000401	.000346	.000305	.000273	.000246	.000225	.000184	.000156	.000135	.000126	.000097	.000082	.000070	.000058	.000050
1.9	.000800	.000724	.000662	.000609	.000564	.000492	.000436	.000392	.000356	.000326	.000300	.000260	.000229	.000205	.000185	.000169	.000138	.000117	.000102	.000090	.000073	.000061	.000053	.000044	.000037
2.0	.000589	.000534	.000488	.000450	.000417	.000364	.000323	.000291	.000264	.000242	.000223	.000193	.000170	.000152	.000138	.000126	.000103	.000088	.000076	.000067	.000054	.000046	.000040	.000033	.000026
2.1	.000430	.000390	.000357	.000330	.000306	.000268	.000238	.000214	.000195	.000178	.000165	.000143	.000126	.000113	.000102	.000093	.000076	.000065	.000056	.000050	.000040	.000034	.000029	.000024	.000021
2.2	.000311	.000283	.000260	.000240	.000223	.000195	.000174	.000156	.000142	.000130	.000120	.000104	.000092	.000083	.000075	.000068	.000056	.000048	.000041	.000037	.000030	.000025	.000022	.000018	.000015
2.3	.000224	.000204	.000187	.000173	.000161	.000141	.000126	.000113	.000103	.000095	.000087	.000076	.000067	.000060	.000054	.000050	.000041	.000035	.000030	.000027	.000022	.000018	.000016	.000013	.000011
2.4	.000160	.000146	.000134	.000124	.000115	.000101	.000090	.000081	.000074	.000068	.000063	.000055	.000048	.000043	.000039	.000036	.000030	.000025	.000022	.000019	.000016	.000013	.000011	.000009	.000008
2.5	.000113	.000103	.000095	.000088	.000082	.000072	.000064	.000058	.000053	.000049	.000045	.000039	.000035	.000031	.000028	.000026	.000021	.000018	.000016	.000014	.000011	.000009	.000008	.000007	.000006
2.6	.000079	.000073	.000067	.000062	.000058	.000051	.000045	.000041	.000037	.000034	.000032	.000028	.000025	.000022	.000020	.000018	.000015	.000013	.000011	.000010	.000008	.000007	.000006	.000005	.000004
2.7	.000055	.000051	.000047	.000043	.000040	.000036	.000032	.000029	.000026	.000024	.000022	.000020	.000017	.000016	.000014	.000013	.000011	.000009	.000008	.000007	.000006	.000005	.000004	.000003	.000003
2.8	.000038	.000035	.000032	.000030	.000028	.000025	.000022	.000020	.000018	.000017	.000016	.000014	.000012	.000011	.000010	.000009	.000007	.000006	.000005	.000005	.000004	.000003	.000003	.000002	.000002
2.9	.000026	.000024	.000022	.000021	.000019	.000017	.000015	.000014	.000013	.000012	.000011	.000009	.000008	.000007	.000007	.000006	.000005	.000004	.000004	.000003	.000003	.000002	.000002	.000002	.000001
3.0	.000018	.000016	.000015	.000014	.000013	.000012	.000010	.000009	.000009	.000008	.000007	.000006	.000006	.000005	.000005	.000004	.000004	.000003	.000003	.000002	.000002	.000002	.000001	.000001	.000001

A NOTE ON THE SUM OF A LINEAR AND LINEAR-FRACTIONAL FUNCTION

Siegfried Schaible

University of Cologne
Cologne, W. Germany

ABSTRACT

The sum of a linear and linear-fractional function is investigated in terms of quasi-convexity and quasi-concavity. From this we obtain some insight into the nature of local optima of these functions useful in algorithms.

Consider the optimization problem

$$\sup \{q(x)=a^T x+b^T x/c^T x \mid x \in S\},\ S\subseteq R^n \text{ convex, } c^T x>0 \tag{1}$$

which arises when a compromise between absolute and relative terms is to be maximized [11]. From a theoretical as well as algorithmic point of view, it is important to get some insight into the nature of local optima of (1). For linear programs ($b=0$) and linear fractional programs ($a=0$), we know that (A1) a local maximum is a global maximum, (A2) a local maximum is attained at an extreme point of S [2]. What are the conditions such that at least one of these assertions is true for the more general problem (1)? Recall that (1) can often be solved by a convex programming procedure if (A1) holds, and a simplex-like procedure can be applied if (A2) is true (S polyhedral) [5].

As known from the theory of generalized convex programming [5], (A1) is essentially equivalent to $q(x)$ being quasi-concave (*qcv*), and (A2) is essentially equivalent to $q(x)$ being quasi-convex (*qcx*) on S. We therefore investigate $q(x)$ in terms of quasi-concavity and quasi-convexity.

It is assumed that $a\neq 0$, $b\neq 0$, and b, c are linearly independent. We need consider only the following cases:

I. a, b, c are linearly independent.
II. a, b are linearly dependent, i.e. $a=\mu b$ ($\mu\neq 0$).
III. a, c are linearly dependent, i.e. $a=\lambda c$ ($\lambda\neq 0$).

By an affine transformation of variables, $y=Ax$, $q(x)$ reduces to

I. $\bar{q}(y)=y_1+y_2/y_3$
II. $\bar{q}(y)=\mu y_2+y_2/y_3$
III. $\bar{q}(y)=\lambda y_3+y_2/y_3$

Such a transformation does not affect quasi-concavity (quasi-convexity) [7]. It can be shown:

PROPOSITION: We have in the above mentioned cases

I: $q(x)$ is neither *qcv* nor *qcx* on an n-dimensional convex set S,

II: $\mu>0$: $q(x)$ is *qcv* on $\{x\epsilon S|a^Tx\leq 0\}$ and *qcx* on $\{x\epsilon S|a^Tx\geq 0\}$, $\mu<0$: $q(x)$ is *qcv* on $\{x\epsilon S|a^Tx\leq 0,\ c^Tx\leq -\mu\}$ and $\{x\epsilon S|a^Tx\geq 0,\ c^Tx\geq -\mu\}$, and *qcx* on $\{x\epsilon S|a^Tx\geq 0,\ c^Tx\leq -\mu\}$ and $\{x\epsilon S|a^Tx\leq 0,\ c^Tx\geq -\mu\}$,

III: $q(x)$ is *qcv* on S if $\lambda<0$, and *qcx* on S if $\lambda>0$.

PROOF: I. Let $\tilde{y}$ be an interior point of $A(S)$, and consider $\bar{q}(y)=y_1+y_2/y_3$ on the intersection I of $A(S)$ with the hyperplane $y_1=-y_3+(\tilde{y}_1+\tilde{y}_3)$. Since the set $\{y\epsilon I|\bar{q}(y)\leq\alpha\}$ for $\alpha=\bar{q}(\tilde{y})$ is not convex, $\bar{q}(y)$ is not *qcx* on I [4]. Then $\bar{q}(y)$ is not *qcx* on $A(S)$. Since $-q(x)$ is of the same form as $q(x)$, $q(x)$ is not *qcv* on S.

II. Here $\bar{q}(y)=y_2/k(y_3)$, where $k(y_3)=y_3/(y_3+\mu)$. Since $k''(y_3)=-2\mu/(y_3+\mu)^3$, $k(y_3)$ is either concave or convex depending on the sign of μ and $y_3+\mu$. Thus, $\bar{q}(y)$ is a quotient of a linear and a concave or convex function. Hence, it is *qcv* or *qcx*, respectively [4].

III. Here $\bar{q}(y)=(\lambda y_3^2+y_2)/y_3$. For a concave (convex) numerator the quotient is *qcv* (*qcx*) [4].

From the Proposition we see:

CASE I: (A1) and (A2) are not true in general (for examples, see Ref. [3]).

CASE II, $\mu>0$ and case III:

(2) (A1) is true, if $a^Tx\leqslant 0$ on S,

(3) (A2) is true, if $a^Tx\geqslant 0$ on S.

CASE II, $\mu<0$: the assertions (2), (3) hold only under more restrictive assumptions (see Proposition).

We see that for a rather limited class of problems (1), local optima have one of the properties (A1), (A2). In Ref. [11] it was stated that (2) and (3) are valid for all problems (1). In particular, the simplex-like procedure in Ref. [11] can be applied only under more restrictive assumptions. If these are not fulfilled, a method by Ritter [6] may be used.

REMARK: As seen above, in Case II and III $q(x)$ can often be written as a quotient of a concave and convex function. Here, an extension of Charnes-Cooper's variable transformation [1] relates (1) to a convex program [8]. Then, duality relations are obtainable for (1) [9, 10], providing other means of solving this problem.

REFERENCES

[1] Charnes, A., and W. W. Cooper, "Programming with Linear Fractional Functionals," Naval Research Logistics Quarterly *9*, 181–186 (1962).

[2] Dinkelbach, W., "Die Maximierung eines Quotienten zweier linearer Funktionen unter linearen Nebenbedingungen," Zeitschrift für Wahrscheinlichkeitstheorie u. verw. Gebiete *1*, 141–145 (1962).

[3] Hirche, J., "Zur Extremwertannahme und Dualität bei Optimierungsproblemen mit Linearem und Gebrochen-Linearem Zielfunktionsanteil," Zeitschrift für Angew. Math. u. Mech. *55*, 184–185 (1975).

[4] Mangasarian, O. L., *Nonlinear Programming*, (McGraw-Hill Book Co., New York 1969).

[5] Martos, B., "The Direct Power of Adjacent Vertex Programming Methods," Management Science *12*, 241–252 (1965).

[6] Ritter, K., "A Parametric Method for Solving Certain Nonconcave Maximization Problems," Journal of Computer and System Sciences *1*, 44–54 (1967).

[7] Schaible, S., "Quasi-convex Optimization in General Real Linear Spaces," Zeitschrift für Operations Research *16*, 205–213 (1972).

[8] Schaible, S., "Parameter-free Convex Equivalent and Dual Programs of Fractional Programming Problems," Zeitschrift für Operations Research *18*, 187–196 (1974).

[9] Schaible, S., "Fractional Programming. I, Duality," Management Science *22*, 858–867 (1976).

[10] Schaible, S., "Duality in Fractional Programming: A Unified Approach," Operations Research *24*, 452–461 (1976).

[11] Teterev, A. G., "On a Generalization of Linear and Piece-wise Linear Programming (Russ.)," Ekon. i Mat. Met. *5*, 440–447 (1969); English translation in Matekon, 246–259 (1970).

COMMUNICATION AND CORRECTION

In my paper, "State-Dependent Gap Acceptance" (December 1976 NRLQ), there is an error in the derivation of the total delay distribution on page 654. It should have been noted that the random sequence $\{T_i\}$ of gaps faced, and the number $(M+1)$ of gaps needed for merging, are *dependent* and not independent stochastic variables. Therefore, the formula for the delay CDF $B_D(t)$ must be recalculated for this synchronous model as

$$B_D(t)=\Pr\{\text{delay } D\leq t\}$$

$$=\sum_{m=0}^{\infty}\Pr\{\text{wait is } m \text{ gaps}\}\cdot\Pr\{D\leq t|M=m\}.$$

Given that it took $(m+1)$ gaps for the merge, we immediately know that every one of the first m gaps faced was less than the critical gap. Hence, the CDF associated with each of these m gaps (say $G(t)$) is the underlying gap distribution truncated at T_0; so $G(t)=F(t)/(1-p)$ for $0<t<T_0$. Thus,

$$B_D(t)=\sum_{m=0}^{\infty}p_m G^{*(m)}(t).$$

Since complete independence has been removed from the random sum D_n, we must rederive an expression for its mean. First, we note from the revised expression for $B_D(t)$ that now

$$E[D]=\sum_{m=0}^{\infty}p_m(m\,E[T|T<T_0])=E[M]\,E[T|T<T_0].$$

All other results in the paper carry through with $G(t)$ and $E[T|T<T_0]$ replacing $F(t)$ and $E[T]$, respectively.

The major new formulae are thus

$$E[D_n]=\frac{1-p_n}{p_n}\,E[T|T<T_0^{(n)}]=\frac{\int_0^{T_0^{(n)}}t\,dF(t)}{p_n}=\frac{\int_0^{T_0^{(n)}}t\,dF(t)}{F(T_0^{(n)})}$$

and

$$T_0^{(n)}=G^{-1}\left[\frac{E[T|T<T_0^{(n)}]}{t_n+E[T|T<T_0^{(n)}]}\right],$$

from which we must now obtain $T_0^{(n)}$.

This leads, then, to revised critical gaps $(T_0^{(1)}, T_0)$ in the illustrative example. There,

(1) $$T_0^{(1)}=G^{-1}\left[\frac{E[T|T<T_0^{(1)}]}{\rho_1+E[T|T<T_0^{(1)}]}\right]$$

and

(2) $$T_0=G^{-1}\left[\frac{E[T|T<T_0]}{\rho+E[T|T<T_0]}\right].$$

But the gap *CDF* is assumed to be

$$F(t)=1-e^{-t}.$$

Thus

$$G(t)=\frac{1-e^{-t}}{1-e^{-T_0^{(1)}}},$$

and

$$G^{-1}(u)=-\ln\left[1-e^{-T_0^{(1)}}\right]. \quad (3)$$

Equations (1), (2), and (3) are then combined to give one nonlinear equation for $T_0^{(1)}$ and another for T_0. These two equations are individually solved approximately to find the final values of the two critical gaps.

Carl M. Harris
Syracuse University
Syracuse, N.Y.

INDEX TO VOLUME 24

INDEX TO VOLUME 24—Continued

INDEX TO VOLUME 24—Continued

INDEX TO VOLUME 24—Continued

☆ U. S. GOVERNMENT PRINTING OFFICE : 1978 261-252/x-1

INFORMATION FOR CONTRIBUTORS

The NAVAL RESEARCH LOGISTICS QUARTERLY is devoted to the dissemination of scientific information in logistics and will publish research and expository papers, including those in certain areas of mathematics, statistics, and economics, relevant to the over-all effort to improve the efficiency and effectiveness of logistics operations.

Manuscripts and other items for publication should be sent to The Managing Editor, NAVAL RESEARCH LOGISTICS QUARTERLY, Office of Naval Research, Arlington, Va. 22217. Each manuscript which is considered to be suitable material for the QUARTERLY is sent to one or more referees.

Manuscripts submitted for publication should be typewritten, double-spaced, and the author should retain a copy. Refereeing may be expedited if an extra copy of the manuscript is submitted with the original.

A short abstract (not over 400 words) should accompany each manuscript. This will appear at the head of the published paper in the QUARTERLY.

There is no authorization for compensation to authors for papers which have been accepted for publication. Authors will receive 250 reprints of their published papers.

Readers are invited to submit to the Managing Editor items of general interest in the field of logistics, for possible publication in the NEWS AND MEMORANDA or NOTES sections of the QUARTERLY.

NAVAL RESEARCH
LOGISTICS
QUARTERLY

DECEMBER 1977
VOL. 24, NO. 4
NAVSO P-1278

CONTENTS

ARTICLES

OFFICE OF NAVAL RESEARCH
Arlington, Va. 22217